NASA SP-69

SPACE TECHNOLOGY

Volume V

TELECOMMUNICATIONS

J. J. STIFFLER
Jet Propulsion Laboratory

Scientific and Technical Information Division 1966
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Washington, D.C.

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C., 20402 - Price 55 cents

Preface

This is the fifth of a series of publications prepared as notes for a course in the Space Technology Summer Institute, given by the California Institute of Technology (Caltech) in cooperation with the Jet Propulsion Laboratory (JPL) from June 19 to July 31, 1964. The Institute—directed by E. E. Sechler, professor of aeronautics at Caltech—was sponsored by the National Aeronautics and Space Administration under Grant No. NsG–598 and was taught by engineers from industry, from Caltech, and from JPL. It is planned that the complete set will consist of—

Volume I	*Spacecraft Systems* by L. H. Abraham DOUGLAS AIRCRAFT CO., INC.
Volume II	*Spacecraft Mechnical Engineering* by James L. Adams, JPL
Volume III	*Spacecraft Propulsion* by F. E. Marble, CALTECH
Volume IV	*Spacecraft Guidance and Control* by J. R. Scull, JPL
Volume V	*Telecommunications* by J. J. Stiffler, JPL

Contents

Introduction

A SPACE MISSION is no more successful than its telemetry. For it is only through the telemetry received from the vehicle that we are able to determine how it is performing, what the engineering and scientific measurements are yielding, and, with the possible exception of large satellites in near-earth orbits, whether the vehicle is even there or not.

Wireless telemetry has been a reality for about 70 years now, and has been used extensively for 40 or 50 years. In that time radio, television, and radar techniques have been developed to extremely high levels of performance. Commercial radio and television has become a way of life; commercial aviation depends upon radio and radar, and even amateurs operate their own radio transmitters. Why, then, should there be any problems associated with telemetry in the space age? It would seem that all the efforts should be devoted to the less-tried engineering pursuits, such as rocket design. Telemetry, albeit important, is well understood.

A little reflection, however, suggests a number of significant differences between space telemetry and surface telemetry. The most striking of these lies in the distances involved. Clearly, until recently, no attempt was ever made to transmit information more than about 12 000 miles, since no two points on earth are separated by more than that distance. (Even communication at distances of more than a few thousand miles depended upon rather unpredictable meteorological phenomena and could not be relied upon.) Yet the nearest neighbor to the earth is about 20 times farther away than that and the planet nearest to the earth more than 2000 times as distant. Since the power at the receiver is inversely proportional to the square of the distance between it and the transmitter, the power received from a transmitter on Venus would be only one four-millionth as great as the power received from the same transmitter placed on the opposite side of the earth.

The second major difference between space telemetry and surface telemetry rests in the constraints placed on the transmitter and receiver in the space vehicle. In contrast to commercial radio and television, in which the transmitter can be large and complex whereas the receivers must be kept small and inexpensive, the transmitter in the spacecraft-to-ground telemetry link is limited by weight and reliability constraints, while the receiver, on the ground, is relatively unconstrained. It is clearly not possible, or at least not practical, to include, as part of a spacecraft, a 50 000-watt transmitter complete with a steam turbine to generate the power and an 80-foot antenna to transmit it. The equipment must be kept as small and as reliable as possible and must be fully automatic; the power required can be no more than that supplied by the generators and batteries on board which, in turn, must be kept within reasonable weight limitations. Thus, in practice, the transmitted power is limited to a few watts.

The receiver, however, may reasonably involve large antennas, a crew of operators, and even a digital computer. The ground-to-vehicle link is more conventional in its constraints, although the reliability demands and the operating conditions to which the receiver is subject are, of course, considerably more severe than in the usual situation.

Granting, then, that space telemetry does, indeed, pose new and unique problems, what are the means which have been adopted for a solution? They, in general, fall into one of two categories: (1) improved components, and (2) improved data handling and modulation systems.

The discussion of these techniques is the purpose of these notes. We begin by reviewing some mathematical techniques and introducing some fundamental concepts in chapter 1. In chapter 2 we investigate some of the methods whereby the effective signal power at the receiver can be vastly increased through improved component design. We then discuss, in some detail in chapter 3, the somewhat conventional modulation techniques and proceed to investigate the more recent pulse-modulation schemes in chapter 4. Having observed some definite advantages inherent in pulse modulation, we then discuss in chapter 5 the data-handling efficiencies possible when working with pulsed, or sampled, data. Chapter 6 is concerned with the related problems of ranging and, briefly, telemetry synchronization. And, finally, chapter 7 discusses the telemetry systems which have actually been used in the Pioneer and Mariner programs and some of the recent innovations used in the earth-based receiving equipment.

CHAPTER 1

Fundamentals

IN ORDER TO DISCUSS, in any detail, the essentials of modern communications techniques it is necessary first to define some terms and to review some mathematical concepts. It is hoped that this chapter will help to clarify some of those concepts such as bandwidth and noise so fundamental to the theory of telecommunications.

FOURIER SERIES AND FOURIER TRANSFORMS

Let $f(t)$ be a function periodic in time with a period T such that $f(t)=f(t+nT)n=0, \pm 1, \pm 2, \ldots$ and with the property that

$$\int_{-T/2}^{T/2} |f(t)| \, dt < \infty .$$

Then, subject to some rather general conditions, $f(t)$ can be expressed as a Fourier series:

$$f(t)=\frac{a_0}{T}+\frac{2}{T}\sum_{n=1}^{\infty} (a_n \cos \omega_n t + b_n \sin \omega_n t). \tag{1.1}$$

where $\omega_n = 2\pi n/T$. Thus $f(t)$ may be considered to be a weighted sum of sinusoids of frequencies $f_n=\omega_n/2\pi=n/T$. The Fourier coefficient a_0 may be evaluated by integrating both sides of equation (1.1) over one period

$$\int_{-T/2}^{T/2} f(t) \, dt = a_0 \tag{1.2}$$

Similarly a_n and b_n can be evaluated by multiplying both sides of equation (1.1) by cos $\omega_n t$ and sin $\omega_n t$, respectively, and integrating the product over one period:

$$\left.\begin{aligned} a_n &= \int_{-T/2}^{T/2} f(t) \cos \omega_n t \, dt \\ b_n &= \int_{-T/2}^{T/2} f(t) \sin \omega_n t \, dt \end{aligned}\right\} \tag{1.3}$$

While we have equated the function $f(t)$ to its Fourier series, it should be noted that this equality is valid only at the points of continuity of $f(t)$.

Several alternative forms of the Fourier series will be useful. Since

$$\cos \omega_n t = \frac{e^{j\omega_n t}+e^{-j\omega_n t}}{2}$$

and

$$\sin \omega_n t = \frac{e^{j\omega_n t}-e^{-j\omega_n t}}{2j}$$

where $j=\sqrt{-1}$, $f(t)$ may be written

$$\begin{aligned} f(t) &= \frac{a_0}{T}+\frac{1}{T}\sum_{n=1}^{\infty}(a_n-jb_n)e^{j\omega_n t}+\frac{1}{T}\sum_{n=1}^{\infty}(a_n+jb_n)e^{-j\omega_n t} \\ &= \frac{1}{T}\sum_{n=-\infty}^{\infty} c_n e^{j\omega_n t} \end{aligned} \tag{1.4}$$

Here $c_{-n}=c_n^*$, the asterisk designating "the complex conjugate of," $c_n=a_n-jb_n$, and $c_0=a_0$. The last series is referred to as the Fourier exponential series. It is readily verified that

$$c_n=\int_{-T/2}^{T/2} f(t)e^{-j\omega_n t}\,dt \tag{1.5}$$

Finally, noting that

$$a_n \cos \omega_n t + b_n \sin \omega_n t = d_n \cos (\omega_n t+\phi_n)$$

where

$$d_n=(a_n^2+b_n^2)^{1/2}$$

and

$$\phi_n = -\tan^{-1} b_n/a_n$$

it is possible to write $f(t)$ as a series of cosines only:

$$f(t)=\frac{2}{T}\sum_{n=0}^{\infty} d_n \cos(\omega_n t+\phi_n) \tag{1.6}$$

Consider now a function $f(t)$ with a period that is actually infinite.

Then letting $\Delta\omega=\omega_{n+1}-\omega_n=2\pi/T$ and taking the limit of equation (1.4) as $T\longrightarrow\infty$ and $n\longrightarrow\infty$ 'n such a way that $\omega_n=(2\pi n/T)\rightarrow\omega$, we find that

$$f(t)=\lim_{n,T\rightarrow\infty}\frac{1}{2\pi}\sum_{n=-\infty}^{\infty}c_n e^{j\omega_n t}\Delta\omega=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(j\omega)e^{j\omega t}\,d\omega \tag{1.7}$$

where

$$\begin{aligned}F(j\omega)&=\lim_{\substack{n,T\rightarrow\infty\\ \omega_n\rightarrow\omega}}c_n=\lim_{\substack{n,T\rightarrow\infty\\ \omega_n\rightarrow\omega}}\int_{-T/2}^{T/2}f(t)e^{-j\omega_n t}\,dt\\ &=\int_{-\infty}^{\infty}f(t)e^{-j\omega t}\,dt\end{aligned} \tag{1.8}$$

The functions $f(t)$ and $F(j\omega)$ as defined above are called Fourier transform pairs, $F(j\omega)$ being the representation in the frequency domain of the time function $f(t)$. Analogously with the Fourier series, the Fourier transform $F(j\omega)$ of $f(t)$ is defined if

$$\int_{-\infty}^{\infty}|f(t)|\,dt<\infty$$

In addition, again subject to certain generally satisfied conditions on $f(t)$, the relationship

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(j\omega)e^{j\omega t}\,d\omega=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{j\omega t}\int_{-\infty}^{\infty}f(u)e^{-j\omega u}\,du\,d\omega$$

holds at all points of continuity.

As an example, consider the function illustrated in figure 1.1. The Fourier series expansion of the periodic function $f(t)$ is easily determined:

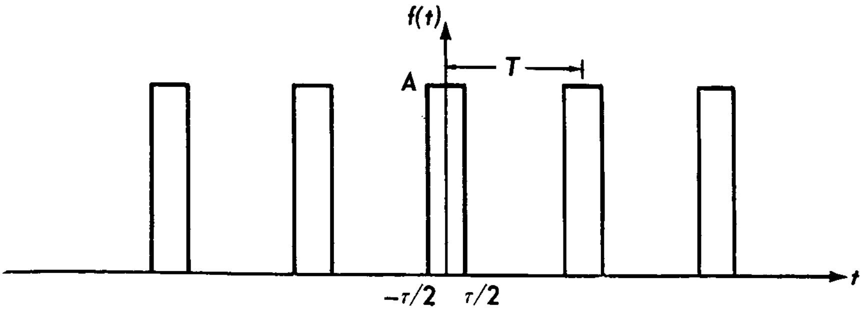


FIGURE 1.1—A periodic function.

$$f(t)=\sum_{n=-\infty}^{\infty}\frac{c_n}{T}\,\mathrm{e}^{j\omega_n t}$$

where (1.9)

$$c_n=A\int_{-\tau/2}^{\tau/2}\mathrm{e}^{-j\omega_n t}\,\mathrm{d}t$$

$$=2A\,\frac{\sin\omega_n\tau/2}{\omega_n}$$

Letting $T\rightarrow\infty$ we find that the Fourier transform of a single pulse of amplitude A and width τ is just

$$F(j\omega)=\lim_{\substack{T\rightarrow\infty\\ \omega_n\rightarrow\omega}}c_n=A\tau\frac{\sin\omega\tau/2}{\omega\tau/2} \tag{1.10}$$

THE DIRAC DELTA FUNCTION

Consider, now, the Fourier transform of the rectangular pulse of the previous section as $\tau\rightarrow 0$ and $A\rightarrow\infty$ in such a way that $A\tau=1$. The pulse becomes an infinitely narrow pulse but with a constant area equal to one. Its Fourier transform becomes

$$F(j\omega)=1$$

The significance of the Fourier transform of this limiting pulse, which we shall call a delta function $\delta(t)$, is highly suspect, since $\delta(t)$ is totally uninteresting except when $t=0$. But because $\delta(t)$ is discontinuous at this point, equation (1.7) is not necessarily satisfied there.

Nevertheless, the delta function has great utilitarian value, and it will be convenient to consider $\delta(t)$ and 1 as Fourier transform pairs. Those who are bothered by this may substitute a pulse of some infinitesimally small but nonzero time duration and of a large but finite amplitude whenever the function $\delta(t)$ occurs in subsequent equations. The results will still be meaningful and true with an arbitrarily small error in most of the manipulations which follow. It will be useful, in fact, in some of the arguments, to consider just such a pulse with width $\Delta\tau$ and amplitude $1/\Delta\tau$. We shall label such a pulse $\delta_{\Delta\tau}(t)$.

An interesting property of the delta function is evidenced when it is integrated. First of all, recall that

$$\int_{-A}^{B} \delta_{\Delta\tau}(t)\ \mathrm{d}t = \int_{-\Delta\tau/2}^{\Delta\tau/2} \delta_{\Delta\tau}(t)\ \mathrm{d}t = 1 \qquad A,B \geq \frac{\Delta\tau}{2}$$

and that consequently

$$\begin{aligned}\int_{-A}^{B} \delta_{\Delta\tau}(t) g(t)\ \mathrm{d}t &= \int_{-\Delta\tau/2}^{\Delta\tau/2} \delta_{\Delta\tau}(t) g(t)\ \mathrm{d}t \\ &\doteq g(0) \int_{-\Delta\tau/2}^{\Delta\tau/2} \delta_{\Delta\tau}(t)\ \mathrm{d}t = g(0)\end{aligned}$$

where the error in replacing $g(t)$, where $-\Delta\tau/2 \leqq t \leqq \Delta\tau/2$, by $g(0)$, of course, goes to zero as $\Delta\tau \rightarrow 0$ and $\delta_{\Delta\tau}(t) \rightarrow \delta(t)$. Thus

$$\left.\begin{aligned}&\int_{-A}^{B} g(t)\delta(t)\ \mathrm{d}t = g(0) && A,B>0 \\ \text{and} \quad &\int_{t_0-A}^{t_0+B} g(t)\delta(t-t_0)\ \mathrm{d}t = g(t_0) && A,B>0\end{aligned}\right\} \qquad (1.11)$$

since $\delta(t-t_0)$ is zero except when $t-t_0=0$ or $t=t_0$. The Fourier transform, by the way, of $\delta(t-t_0)$ is clearly $\mathrm{e}^{-j\omega t_0}$.

By analogy, a delta function can be defined in the frequency domain, $\delta(\omega-\omega_0)$. The corresponding time function then is evidently $f(t) = (1/2\pi)\mathrm{e}^{j\omega_0 t}$. By using delta functions, it is now possible to define the *transform* of a periodic function (although in this case

$$\int_{-\infty}^{\infty} |f(t)|\ \mathrm{d}t = \infty).$$

Since for a periodic function

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{c_n}{T} \mathrm{e}^{j\omega_n t}$$

$$\begin{aligned}\text{then } F(j\omega) &= \int_{-\infty}^{\infty} f(t)\mathrm{e}^{-j\omega t}\ \mathrm{d}t \\ &= \sum_{n=-\infty}^{\infty} \frac{c_n}{T} \int_{-\infty}^{\infty} \mathrm{e}^{-j(\omega-\omega_n)t}\ \mathrm{d}t \\ &= \sum_{n=-\infty}^{\infty} \frac{c_n}{T} 2\pi\delta(\omega-\omega_n) \qquad (1.12)\end{aligned}$$

While in this instance the interpretation of a delta function as a limiting pulse is not meaningful, the term $\delta(\omega-\omega_n)$ can be regarded as a formalism whereby a discrete function $\{c_n\}$ may be writen as a continuous function, $F(j\omega)$.

LINEAR SYSTEMS

Consider the system illustrated in figure 1.2. The time function

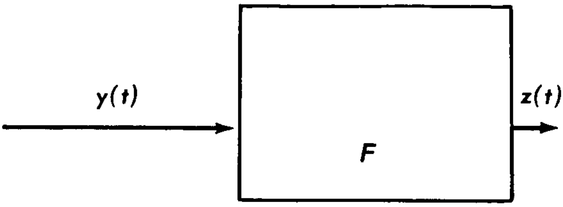


FIGURE 1.2—A linear system.

$y(t)$ is designated as the input and $z(t)$ as the output. The network or *filter* F converts the function $y(t)$ to the function $z(t)$. It is assumed to be *linear;* that is, if $y_1(t)$ produces the output $z_1(t)$ and $y_2(t)$ produces the output $z_2(t)$, then the input $ay_1(t)+by_2(t)$ produces the output $az_1(t)+bz_2(t)$. From this definition it follows that when the input is

$$\sum_{i=0}^{n} a(\tau_i)\Delta\tau_i y(t+\tau_i)$$

the output of F is

$$\sum_{i=0}^{n} a(\tau_i)\Delta\tau_i z(t+\tau_i)$$

and, hence, passing to the limit as $\Delta\tau_i=\tau_i-\tau_{i-1}$ approaches zero while $n\rightarrow\infty$ and $\tau_i\rightarrow\tau$, $\tau_n\rightarrow T$ the input

$$y'(t)=\int_0^T a(\tau)y(t+\tau)\,d\tau$$

produces the output

$$z'(t)=\int_0^T a(\tau)z(t+\tau)\,d\tau$$

Let us define $h(t)$ as the output of the filter F at time t when the input is a delta function at time $t=0$. Then suppose the input is

$$y_\Delta(t)=\sum_{i=-\infty}^{\infty} y(\tau_i)\Delta\tau\delta_{\Delta\tau}(t-\tau_i) \tag{1.13}$$

where $\tau_i=i\Delta\tau$. Such an input is depicted in figure 1.3.

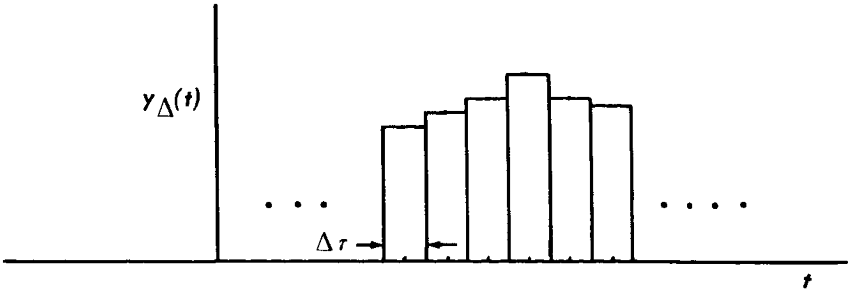


FIGURE 1.3—Example of an input of the form (1.13).

Because of the assumed linearity of the filter, the output is simply

$$\sum_{i=-\infty}^{\infty} y(i\Delta\tau)h_{\Delta\tau}(t-i\Delta\tau)\Delta\tau \tag{1.14}$$

where $h_{\Delta\tau}(t-i\Delta\tau)$ is the response of the filter to a pulse of width $\Delta\tau$ and amplitude $1/\Delta\tau$ occurring at time $t=i\Delta\tau$. Again, taking the limit as $\Delta\tau\to 0$ and $i\to\infty$ in such a way that $i\Delta\tau\to\tau$, the input becomes

$$\int_{-\infty}^{\infty} y(\tau)\delta(t-\tau)\ d\tau=y(t) \tag{1.15}$$

while the output may be written

$$z(t)=\int_{-\infty}^{\infty} y(\tau)h(t-\tau)\ d\tau=\int_{-\infty}^{\infty} y(t-\eta)h(\eta)d\eta \tag{1.16}$$

Thus, the response of the filter to an arbitrary input $y(t)$ can be obtained in terms of the *impulse response* $h(t)$.

There are two practical conditions commonly placed on $h(t)$. The first of these is that the filter F should not be expected to produce an output before it receives an input; that is, $h(t)=0$ when $t<0$. This is the condition that F be a *realizable* filter. We shall not be particularly concerned with realizability here, however. The second condition is that the filter be *stable;* that is, if the input $y(t)$ is bounded, $|y(t)|\leq A$ for all t, then the output must also be bounded. Thus, using equation (1.16)

$$|z(t)|=\left|\int_{-\infty}^{\infty} y(t-\eta)h(\eta)\ d\eta\right|\leq\int_{-\infty}^{\infty}|y(t-\eta)||h(\eta)|\ d\eta\leq A\int_{-\infty}^{\infty}|h(\eta)|\ d\eta \tag{1.17}$$

and if the integral of the absolute value of $h(t)$ is finite, $z(t)$ must be bounded and the filter F is stable. Suppose now that

$$\int_{-\infty}^{\infty} |h(t)|\, \mathrm{d}t = \infty$$

Then, by choosing the input

$$y(t_0-\eta) = \begin{cases} 1 & h(\eta) \geq 0 \\ -1 & h(\eta) < 0 \end{cases}$$

we have demonstrated a bounded input which produces an unbounded output, for

$$z(t_0) = \int_{-\infty}^{\infty} y(t_0-\eta)h(\eta)\, \mathrm{d}\eta = \int_{-\infty}^{\infty} |h(\eta)|\, \mathrm{d}\eta = \infty \tag{1.18}$$

Thus, if the filter is stable $h(t)$ is absolutely integrable. Consequently,

$$|H(j\omega)| \equiv \left|\int_{-\infty}^{\infty} h(t)\mathrm{e}^{-j\omega t}\, \mathrm{d}t\right| \leq \int_{-\infty}^{\infty} |h(t)|\, \mathrm{d}t < \infty$$

and the Fourier transform $H(j\omega)$ of $h(t)$ exists.

Assuming now that $h(t)$, $y(t)$, and $z(t)$ have Fourier transforms and transforming both sides of equation (1.16), we obtain

$$\int_{-\infty}^{\infty} z(t)\mathrm{e}^{-j\omega t}\, \mathrm{d}t = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y(t-\eta)h(\eta)\mathrm{e}^{-j\omega(t-\eta)}\mathrm{e}^{-j\omega\eta}\, \mathrm{d}\eta\, \mathrm{d}t$$

Interchanging the order of integration and defining

$$Y(j\omega) = \int_{-\infty}^{\infty} y(t-\eta)\mathrm{e}^{-j\omega(t-\eta)}\, \mathrm{d}t = \int_{-\infty}^{\infty} y(u)\mathrm{e}^{-j\omega u}\, \mathrm{d}u$$

and

$$Z(j\omega) = \int_{-\infty}^{\infty} z(t)\mathrm{e}^{-j\omega t}\, \mathrm{d}t$$

we have

$$Z(j\omega) = Y(j\omega)H(j\omega) \tag{1.19}$$

The transform $H(j\omega)$ of the impulse response $h(t)$ is referred to as the *transfer function* of the filter F.

Note that if $y(t)$ is periodic, we have from equation (1.4) that

$$Y(j\omega)=\sum_{n=-\infty}^{\infty}\frac{c_n}{T}\,2\pi\delta(\omega-\omega_n)$$

and that, consequently

$$z(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}Z(j\omega)e^{j\omega t}\,d\omega=\sum_{n=-\infty}^{\infty}\frac{c_n}{T}\int_{-\infty}^{\infty}H(j\omega)e^{j\omega t}\delta(\omega-\omega_n)\,d\omega$$

$$=\sum_{n=-\infty}^{\infty}\frac{c_n}{T}H(j\omega_n)e^{j\omega_n t} \tag{1.20}$$

This result can be made more meaningful, perhaps, by considering the response to the input

$$y(t)=\sum_{n=-\infty}^{\infty}\frac{c_n}{T}e^{j\omega_n t}$$

Since F is linear, and since the input is periodic, the output must also be periodic so that

$$z(t)=\sum_{n=-\infty}^{\infty}\frac{c_n'}{T}e^{j\omega_n t}$$

Thus, defining $H(j\omega_n)e^{j\omega_n t}$ to be the response of the filter to a "sinusoid" $e^{j\omega_n t}$ of unit amplitude, $c_n'=c_nH(j\omega_n)$ yields the desired result. That this definition of $H(j\omega)$ is consistent with its previous one as the transform of the impulse response clearly follows because the "transform" of $e^{j\omega_n t}$ is $2\pi\delta(\omega-\omega_n)$. Then, from equation (1.19), when $y(t)=e^{j\omega_n t}$

$$Z(j\omega)=2\pi H(j\omega)\delta(\omega-\omega_n)$$

and

$$z(t)=\int H(j\omega)\delta(\omega-\omega_n)e^{j\omega t}\,d\omega=H(j\omega_n)e^{j\omega_n t}$$

POWER AND ENERGY SPECTRA

Let $y(t)$ be a periodic function

$$y(t)=\sum_{n=-\infty}^{\infty}\frac{c_n}{T}e^{j\omega_n t}\qquad \omega_n=\frac{2\pi n}{T} \tag{1.21}$$

The *average power* in $y(t)$ is defined as

$$P_{\text{ave}}=\frac{1}{T}\int_{-T/2}^{T/2}y^2(t)\,dt \tag{1.22}$$

(If $y(t)$ is a voltage level, for example, P_{ave} represents the average power dissipated by passing $y(t)$ through a 1-ohm resistor). Substituting equation (1.21) into (1.22), we obtain

$$P_{\text{ave}}=\sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\frac{c_n}{T}\frac{c_m}{T}\frac{1}{T}\int_{-T/2}^{T/2} e^{j(2\pi/T)(n+m)t}\,dt$$

Clearly, the integral on the left vanishes except when $m=-n$, in which case it is just T.

Thus

$$P_{\text{ave}}=\sum_{n=-\infty}^{\infty}\left|\frac{c_n}{T}\right|^2 \tag{1.23}$$

where use is made of the fact that $c_{-n}=c_n{}^*$.

Suppose now we pass $y(t)$ through a filter with the property that only the mth component is passed; that is

$$H(j\omega_n)=\begin{cases}1 & n=\pm m \quad \text{(see footnote 1)}\\ 0 & \text{Otherwise}\end{cases}$$

The output of the filter is then

$$\begin{aligned} z(t) &=\frac{c_m}{T}H(j\omega_m)e^{j\omega_m t}+\frac{c_{-m}}{T}H(-j\omega_m)e^{-j\omega_m t}\\ &=\left[\frac{c_m}{T}e^{j\omega_m t}\right]+\left[\frac{c_m}{T}e^{j\omega_m t}\right]^*\\ &=2\text{Re}\left[\frac{c_m}{T}e^{j\omega_m t}\right]=2\left|\frac{c_m}{T}\right|\cos\left(\omega_m t+\theta_m\right) \end{aligned} \tag{1.24}$$

where $\text{Re}(X)$ designates the real part of X. The average power at the frequency ω_m is thus

$$\frac{4}{T}\int_{-T/2}^{T/2}\left|\frac{c_m}{T}\right|^2\cos^2\left(\omega_m t+\theta_m\right)\,dt=2\left|\frac{c_m}{T}\right|^2 \tag{1.25}$$

[1] Note that $|H(j\omega)|=|H(-j\omega)|$ for any filter with an impulse response which is a real function of time:

$$\begin{aligned} H(j\omega)&=\int_{-\infty}^{\infty}h(t)\cos\omega t\,dt-j\int_{-\infty}^{\infty}h(t)\sin\omega t\,dt\\ &=A(\omega)-jB(\omega)\end{aligned}$$

where $A(\omega)$ and $B(\omega)$ are both real functions of ω. Then since $H(-j\omega)=A(\omega)+jB(\omega)=H^*(j\omega)$, the statement follows.

For the case $m=0$, it is easily verified that the average power is $|c_0/T|^2$. After rewriting equation (1.23)

$$P_{\text{ave}}=\left|\frac{c_0}{T}\right|+2\sum_{n=1}^{\infty}\left|\frac{c_n}{T}\right|^2 \tag{1.26}$$

it is clear that the average power in the function $f(t)$ is just the sum of the average powers in each of its Fourier components.

Consider the case now in which $y(t)$ is not periodic but has a Fourier transform $Y(j\omega)$. Then

$$y(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} Y(j\omega)e^{j\omega t}\,\mathrm{d}\omega \tag{1.27}$$

and

$$P_{\text{ave}}=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} y^2(t)\,\mathrm{d}t \tag{1.28}$$

But if $y(t)$ has a bounded amplitude, $|y(t)|\leq A$,

$$\int_{-\infty}^{\infty} y^2(t)\,\mathrm{d}t=A^2\int_{-\infty}^{\infty}\left(\frac{y(t)}{A}\right)^2\mathrm{d}t\leq A\int_{-\infty}^{\infty}|y(t)|\,\mathrm{d}t$$

and unless this last integral is finite, $Y(j\omega)$ need not exist. Generally, then $P_{\text{ave}}=0$ for functions which have Fourier transforms. It is, nevertheless, interesting to consider the *energy* E_y in the function $y(t)$:

$$E_y=\int_{-\infty}^{\infty} y^2(t)\,\mathrm{d}t \tag{1.29}$$

Substituting equation (1.27) into this expression and changing the order of integration yields

$$\begin{aligned}E_y&=\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Y(j\omega_1)Y(j\omega_2)\int_{-\infty}^{\infty} e^{j(\omega_1+\omega_2)t}\,\mathrm{d}t\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2\\&=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Y(j\omega_1)Y(j\omega_2)\delta(\omega_1+\omega_2)\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2\\&=\frac{1}{2\pi}\int_{-\infty}^{\infty}|Y(j\omega)|^2\,\mathrm{d}\omega\end{aligned} \tag{1.30}$$

Let us, as we did in the case of the periodic function, pass $y(t)$ through a filter with a response

$$H(j\omega)=\left\{\begin{array}{ll} 1 & \omega_0-\frac{\Delta\omega}{2}<\omega<\omega_0+\frac{\Delta\omega}{2} \\ 1 & -\omega_0-\frac{\Delta\omega}{2}<\omega<-\omega_0+\frac{\Delta\omega}{2} \\ 0 & \text{Otherwise} \end{array}\right\} \tag{1.31}$$

Then $Z(j\omega)=Y(j\omega)H(j\omega)$ and

$$\begin{aligned} E_z&=\int_{-\infty}^{\infty} z^2(t)\,dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}|Y(j\omega)H(j\omega)|^2\,d\omega \\ &\doteq\left\{|Y(j\omega_0)|^2+|Y(-j\omega_0)|^2\right\}\frac{\Delta\omega}{2\pi} \end{aligned} \tag{1.32}$$

where $\Delta\omega$ is assumed to be small enough so that $Y(j\omega)\doteq Y(j\omega_0)$ over the intervals in question. Since $|Y(j\omega)|=|Y(-j\omega)|$ and letting $\Delta f=\Delta\omega/2\pi$, we have

$$\epsilon(\omega_0)\equiv\frac{E_z(\omega_0)}{\Delta f}=2|Y(j\omega_0)|^2 \tag{1.33}$$

Thus, equations (1.23) and (1.30) do not only indicate the average power and energy in the functions being investigated, but also indicate the power and energy levels at the various frequency components. For this reason the function defined in equation (1.33) is referred to as the *energy spectral density* of the function $y(t)$. Similarly, using delta functions, it is possible to write a *power spectral density* $P(\omega)$ for a periodic function $y(t)$ as follows:

$$P(\omega)=2\pi\left|\frac{c_0}{T}\right|^2\delta(\omega)+4\pi\sum_{n=1}^{\infty}\left|\frac{c_n}{T}\right|^2\delta(\omega-\omega_n) \tag{1.34}$$

Then

$$P_{\text{ave}}=\frac{1}{2\pi}\int_0^{\infty}P(\omega)\,d\omega=\left|\frac{c_0}{T}\right|^2+2\sum_{n=1}^{\infty}\left|\frac{c_n}{T}\right|^2$$

The energy and power spectral densities of equations (1.33) and (1.34) are called *single-sided* spectral densities since they are both defined only for positive frequencies (both $-\omega$ and ω contributions were combined in one expression) and correspond to the actual power or energy that would be measured at that frequency. The corresponding *two-sided* spectral densities are defined in the obvious way as

$$\epsilon(\omega) = |Y(j\omega)|^2 \tag{1.35}$$

and

$$P(\omega) = \sum_{n=-\infty}^{\infty} 2\pi \left|\frac{c_n}{T}\right|^2 \delta(\omega - \omega_n) \tag{1.36}$$

Note that single- and two-sided densities differ by a factor of 2 except, in the periodic case, at zero frequency.

SPECTRA AND AUTOCORRELATION

Thus far we have considered two kinds of processes: those characterized by periodic functions of time and those characterized by *transient* functions of time (i.e., those which have Fourier transforms and consequently must decay in time). From the communications point of view, both types of processes are uninteresting; both are deterministic and hence completely predictable. Information is transmitted only when the receiver does not know exactly what to expect from the sender. So far as the receiver is concerned, the signal is a *randcm* process. An important characterization of a random process $y(t)$ and the one that will be most useful to us here is that afforded by the probability density function $p(y_t)$. This density function has the property that the probability that $a \leq y_t \leq b$, where y_t is the value of $y(t)$ at a particular instant of time t is determined from the integral

$$Pr\{a \leq y_t \leq b\} = \int_a^b p(y_t)\, dy_t$$

For the kinds of signals with which we shall be concerned, $p(y_t)$ will not be a function of the time t. In fact, the random processes $y(t)$ which we will encounter in these notes will fall into an even more restricted class called *stationary* random processes which, among other things, have this property that $p(y_t)$ is not a function of time.

Presumably the average value of the absolute amplitude of a stationary random process $y(t)$ is not zero. Accordingly, $\int_{-\infty}^{\infty} |y(t)|\, dt = \infty$ and the Fourier transform of $y(t)$ does not necessarily exist. (If $\int_{-\infty}^{\infty} |y(t)|\, dt$ were finite, the average value of $|y(t)|$ would be zero, and since the distribution of $y(t)$ does not change with time, $y(t)$ must necessarily be everywhere zero.) To gain insight into the relationship between the random function $y(t)$ and its power spectrum, it is useful to digress and consider the spectrum of periodic and transient functions from another point of view.

Let $y(t)$ be a periodic function.

Then

$$y(t)=\sum_{n=-\infty}^{\infty}\frac{c_n}{T}\,e^{j\omega_n t} \tag{1.37}$$

Define the *autocorrelation function* $\phi_y(\tau)$ of $y(t)$ as

$$\phi_y(\tau)=\frac{1}{T}\int_{-T/2}^{T/2} y(t)y(t+\tau)\,dt \tag{1.38}$$

Note that since $y(t)$ is periodic, $y(t)=y(t+mT)$, where $m=1, 2, \ldots$, $\phi_y(\tau)$ is periodic with the same period, for

$$\begin{aligned}\phi_y(\tau+mT)&=\frac{1}{T}\int_{-T/2}^{T/2} y(t)y(t+\tau+mT)\,dt\\ &=\frac{1}{T}\int_{-T/2}^{T/2} y(t)y(t+\tau)\,dt=\phi_y(\tau)\end{aligned}$$

Intuitively, $\phi_y(\tau)$ is a measure of how much $y(t)$ changes in τ seconds. Note, in particular, that

$$\phi_y(0)>|\phi_y(\tau)| \qquad \text{for } \tau\neq nT,\ n=0,\ \pm1,\ \pm2,\ \ldots \tag{1.39}$$

unless $y(t)$ is a constant, because if $y(t)\neq$constant and $\tau\neq nT$, then

$$\begin{aligned}0&<\int_{-T/2}^{T/2}[y(t)\pm y(t+\tau)]^2\,dt\\ &=\int_{-T/2}^{T/2}y^2(t)\,dt+\int_{-T/2}^{T/2}y^2(t+\tau)\,dt\pm2\int_{-T/2}^{T/2}y(t)y(t+\tau)\,dt\end{aligned}$$

But

$$\int_{-T/2}^{T/2}y^2(t)\,dt=\int_{-T/2}^{T/2}y^2(t+\tau)\,dt=T\phi_y(0)$$

and, hence

$$2T[\phi(0)\pm\phi(\tau)]>0 \qquad \tau\neq nT$$

from which the above statement follows. Thus, unless it is a constant function, $y(t)$ "looks" more like itself than any time shift of itself, substantiating the intuitive interpretation just mentioned. Note, too, that

$$\begin{aligned}\phi_y(-\tau)&=\frac{1}{T}\int_{-T/2}^{T/2}y(t)y(t-\tau)\,dt\\ &=\frac{1}{T}\int_{(-T/2)-\tau}^{(T/2)-\tau}y(u+\tau)y(u)\,du\end{aligned}$$

and because y(t) is periodic with period T

$$\phi_y(-\tau)=\frac{1}{T}\int_{-T/2}^{T/2} y(u+\tau)y(u)\ \mathrm{d}u=\phi_y(\tau)$$

and $\phi_y(\tau)$ is an even function of τ.

Now, since $\phi_y(\tau)$ is periodic with period T, it can be expanded in a Fourier series:

$$\phi_y(\tau)=\frac{1}{T}\sum_{n=-\infty}^{\infty}\Phi_y(\omega_n)\mathrm{e}^{j\omega_n\tau} \qquad \omega_n=\frac{2\pi n}{T} \tag{1.40}$$

where $\Phi_y(\omega_n)$ are the Fourier coefficients of $\phi_y(\tau)$:

$$\Phi_y(\omega_n)=\int_{-T/2}^{T/2}\phi_y(\tau)\mathrm{e}^{-j\omega_n\tau}\ \mathrm{d}\tau \tag{1.41}$$

Substituting from equation (1.38)

$$\begin{aligned}\Phi_y(\omega_n)&=\frac{1}{T}\int_{-T/2}^{T/2}\int_{-T/2}^{T/2} y(t)y(t+\tau)\mathrm{e}^{-j\omega_n\tau}\ \mathrm{d}t\ \mathrm{d}\tau\\ &=\frac{1}{T}\int_{-T/2}^{T/2} y(t)\mathrm{e}^{j\omega_n t}\int_{-T/2}^{T/2} y(t+\tau)\mathrm{e}^{-j\omega_n(t+\tau)}\ \mathrm{d}\tau\ \mathrm{d}t\end{aligned} \tag{1.42}$$

Because $y(t)$ and $\mathrm{e}^{j\omega_n t}$ are both periodic with period T

$$\begin{aligned}\int_{-T/2}^{T/2} y(t+\tau)\mathrm{e}^{-j\omega_n(t+\tau)}\ \mathrm{d}\tau&=\int_{(-T/2)+\tau}^{(T/2)+\tau} y(u)\mathrm{e}^{-j\omega_n u}\ \mathrm{d}u\\ &=\int_{-T/2}^{T/2} y(u)\mathrm{e}^{-j\omega_n u}\ \mathrm{d}u=c_n\end{aligned}$$

and

$$\Phi_y(\omega_n)=\frac{c_n}{T}\int_{-T/2}^{T/2} y(t)\mathrm{e}^{j\omega_n t}\ \mathrm{d}t=\frac{1}{T}|c_n|^2 \tag{1.43}$$

Consequently, the coefficients of the Fourier expansion of the autocorrelation function $\phi_y(\tau)$ are just the terms of the power spectrum of $y(t)$:

$$\begin{aligned}\phi_y(\tau)&=\sum_{n=-\infty}^{\infty}\frac{1}{T}\Phi_y(\omega_n)\mathrm{e}^{j\omega_n\tau}\\ &=\sum_{n=-\infty}^{\infty}\left|\frac{c_n}{T}\right|^2\mathrm{e}^{j\omega_n\tau}\end{aligned} \tag{1.44}$$

Now consider a transient function

$$y(t)=\int_{-\infty}^{\infty} Y(j\omega)\mathrm{e}^{j\omega t}\frac{\mathrm{d}\omega}{2\pi} \qquad (1.45)$$

and define the autocorrelation function

$$\phi_y(\tau)=\int_{-\infty}^{\infty} y(t)y(t+\tau)\,\mathrm{d}t \qquad (1.46)$$

(Note that this definition is different from that in the periodic case in that it is not normalized by the period, which in this case is infinite. The definition of the autocorrelation function for random functions will again be normalized. The definition

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} y(t)y(t+\tau)\,\mathrm{d}t$$

is not used in this case because this integral is in general identically zero for transient functions.)

Substituting equation (1.45) into the defining expression for $\phi_y(\tau)$ yields the result that

$$\begin{aligned}\phi_y(\tau)&=\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Y(j\omega_1)Y(j\omega_2)\mathrm{e}^{j(\omega_1+\omega_2)t}\mathrm{e}^{j\omega_2\tau}\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2\,\mathrm{d}t\\ &=\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Y(j\omega_1)Y(j\omega_2)\mathrm{e}^{j\omega_2\tau}\int_{-\infty}^{\infty}\mathrm{e}^{j(\omega_1+\omega_2)t}\,\mathrm{d}t\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2\\ &=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Y(j\omega_1)Y(j\omega_2)\mathrm{e}^{j\omega_2\tau}\delta(\omega_1+\omega_2)\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2\\ &=\frac{1}{2\pi}\int_{-\infty}^{\infty}|Y(j\omega)|^2\,\mathrm{e}^{j\omega\tau}\,\mathrm{d}\omega \qquad (1.47)\end{aligned}$$

Thus $\phi_y(\tau)$ and $\Phi_y(\omega)=|Y(j\omega)|^2=\epsilon(\omega)$ form Fourier transform pairs. (The reader may verify that $\phi_y(\tau)$ is an even function of τ and that $\phi_y(0)>|\phi_y(\tau)|$, $y(t)\not\equiv$ constant, in this case too.)

For both transient functions and periodic functions then, it is seen that the Fourier transform of the autocorrelation function $\phi_y(\tau)$ is related to the spectral density of the time function $y(t)$. It, therefore, should not be surprising that the same relationship holds for random functions. The Weiner-Khintchine theorem states, in fact, that if $y(t)$ is a random process and if the autocorrelation function is defined

$$\phi_y(\tau)\equiv\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} y(t)y(t+\tau)\,\mathrm{d}t \qquad (1.48)$$

then its Fourier transform

$$\Phi_y(\tau) = \int_{-\infty}^{\infty} \phi_y(\tau) e^{-j\omega\tau}\,d\tau \tag{1.49}$$

is the power spectral density of the process $y(t)$. The proof of this theorem is somewhat involved and is not attempted in this report. However, the importance of the theorem should be immediately apparent. As argued, the random process $f(t)$ will not generally have a Fourier transform. Yet random processes are most useful in characterizing the signals involved in a communication system. The autocorrelation functions of these random processes can often be determined directly from the mathematical description of the system under investigation. Because of this and because of the Wiener-Khintchine relationship, among other things, the autocorrelation function is a powerful analytical tool. While it is beyond the scope of this report to investigate autocorrelation functions in general, two extremely important cases, which will prove to be useful in later analyses, are considered.

The first is the situation in which the power spectral density is constant, independent of ω; $\Phi_y(\omega) = N_0/2$. (This spectrum is called *white* since, analogous with the color white, it contains all frequencies.)

Then

$$\phi_y(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_y(\omega) e^{j\omega\tau}\,d\omega$$

$$= \frac{N_0}{4\pi}\int_{-\infty}^{\infty} e^{j\omega\tau}\,d\omega = \frac{N_0}{2}\delta(\tau)$$

and

$$\phi_y(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} y(t)y(t+\tau)\,dt = \begin{cases} 0 & \tau \neq 0 \\ \infty & \tau = 0 \end{cases}$$

Thus, if the spectral density of the process $y(t)$ is constant for all frequencies, the autocorrelation function of $y(t)$ is identically zero for all values of $\tau \neq 0$; $y(t)$ and $y(t+\tau)$ are said to be uncorrelated. Physically, this means that $y(t)$seems to have no relationship to the value of $y(t+\tau)$. Note, however, that since $\Phi_y(\omega) = N_0/2$

$$P_{\text{ave}} = \int_{-\infty}^{\infty} \Phi_y(\omega)\frac{d\omega}{2\pi} = \phi_y(0) = \infty$$

Consequently, no physical process can have a white spectrum. Nevertheless, white processes are convenient fictions and, in fact, are extremely good approximations to the very common situation in which the spectral density is constant over a bandwidth much greater than the bandwidth of the system under investigation. Suppose that a white process with the spectral density $N_0/2$ were passed through a system which could be represented by a transfer function

$$H(j\omega) = \begin{cases} 1 & |\omega| < 2\pi W \\ 0 & \text{Otherwise} \end{cases}$$

Then the average power at the output of this system due to the noise would be

$$P_{\text{ave}} = \frac{N_0}{4_\pi} \int_{-2\pi W}^{2\pi W} \mathrm{d}W = N_0 W$$

a finite quantity even though the input power were supposedly infinite. The fact that the input spectrum is constant only out to a frequency of say $10W$ does not alter the fact that so far as the system is concerned, it is effectively white. In the time domain, the autocorrelation function of an "almost" white process after being passed through the system with the transfer function $H(j\omega)$ defined previously is given by

$$\phi_y(\tau) = \frac{N_0}{2} \int_{-2\pi W}^{2\pi W} e^{j\omega\tau} \frac{\mathrm{d}\omega}{2\pi} = N_0 W \frac{\sin 2\pi W\tau}{2\pi W\tau}$$

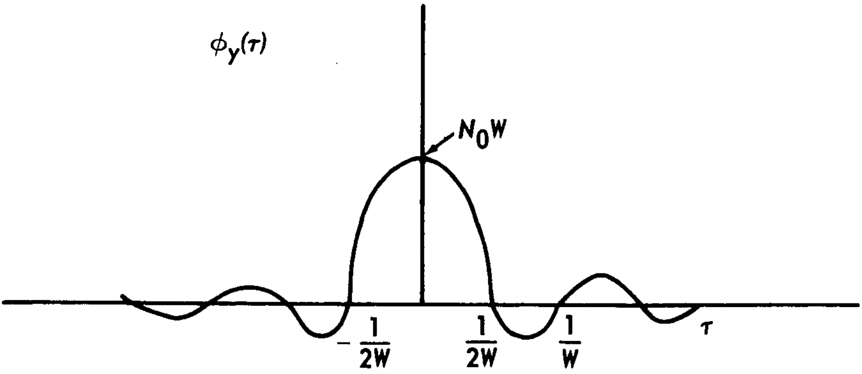


FIGURE 1.4—The (sin x)/x function.

and is graphically illustrated in figure 1.4. Thus $\phi_y(\tau)$, instead of being a delta function of amplitude $N_0/2$, is essentially of pulse width $1/2W$ and amplitude $(N_0/2)2W$, for some large but finite value of W.

The second situation which we will find useful to investigate is that in which the function $y(t)$ is of the form shown in figure 1.5.

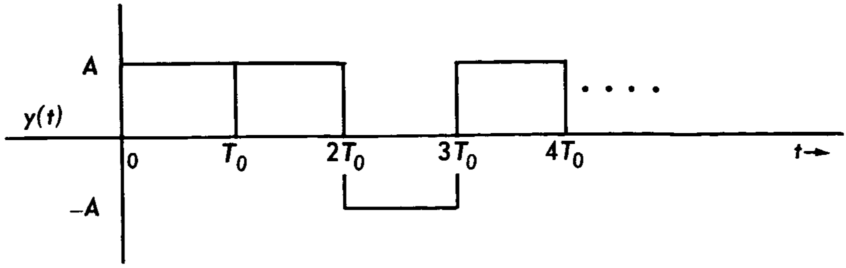


FIGURE 1.5—A random pulse train.

The function $y(t)$ can assume only the values A and $-A$. Each pulse lasts exactly T_0 seconds, at the end of which time $y(t)$ may remain as it was or switch to the other amplitude, with each alternative occurring on an average of exactly one-half the time. Clearly

$$\phi(0)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} y^2(t)\ dt=A^2$$

To determine

$$\phi_y(T_0)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} y(t)y(t+T_0)\ dt$$

observe that $y(t)$ and $y(t+T_0)$ are independent for any value of t, with both functions assuming the values A and $-A$ equally often. Thus

$$y(t)y(t+T_0)=\begin{cases} A\cdot A=A^2 & \text{¼ of the time}\\ A\cdot(-A)=-A^2 & \text{¼ of the time}\\ -A\cdot(A)=-A^2 & \text{¼ of the time}\\ (-A)(-A)=A^2 & \text{¼ of the time}\end{cases}$$

and the average value of this product is zero. Thus $\phi_y(T_0)=0$. Similarly $\phi_y(nT_0)=0$, $n=\pm1, \pm2, \pm3, \ldots$. Now consider $\phi_y(\tau)$, $0<\tau<T_0$. This situation is most easily explained by referring to figure 1.6.

We see that the ith pulse is multiplied by itself for $(T_0-\tau)/T_0$ percent of the time and by the $i+1$st pulse τ/T_0 percent of the time. Thus

$$\begin{aligned}\phi_y(\tau)&=\left(\frac{T_0-\tau}{T_0}\right)\phi_y(0)+\frac{\tau}{T_0}\phi_y(T_0)\\ &=\left(1-\frac{\tau}{T_0}\right)A^2 \qquad 0<\tau<T_0\end{aligned}$$

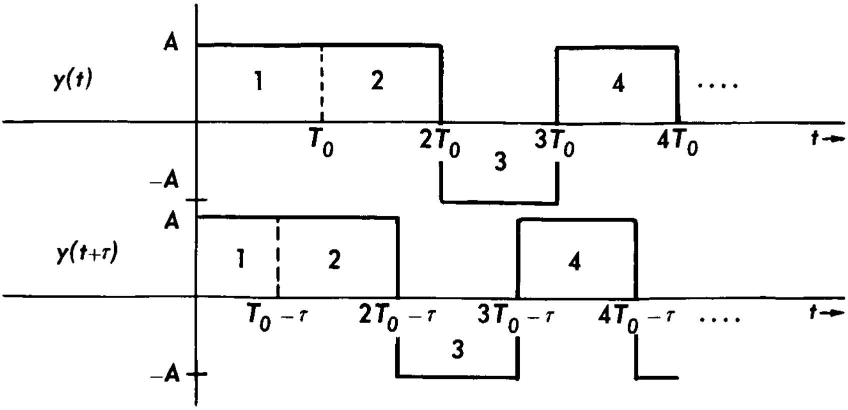


FIGURE 1.6—A random pulse train and its translation in time.

Clearly, the same argument establishes that

$$\begin{aligned}\phi_y(\tau) &= \left(1-\frac{\tau}{T_0}\right)\phi_y(nT_0)+\frac{\tau}{T_0}\phi_y[(n+1)T_0]\\ &=0 \qquad nT_0<\tau<(n+1)T_0 \qquad (n=1,\ 2,\ \ldots)\end{aligned}$$

By symmetry then

$$\phi_y(\tau)=\begin{cases}\left(1-\frac{|\tau|}{T_0}\right)A^2 & |\tau|<T_0\\ 0 & \text{Otherwise}\end{cases}$$

Thus

$$\begin{aligned}\Phi_y(\omega) &= \int_{-\infty}^{\infty}\phi_y(\tau)e^{-j\omega\tau}\,d\tau\\ &=A^2\int_0^{T_0}\left(1-\frac{\tau}{T_0}\right)e^{-j\omega\tau}\,d\tau+A^2\int_{-T_0}^{0}\left(1+\frac{\tau}{T_0}\right)e^{-j\omega\tau}\,d\tau\\ &=A^2T_0\,\frac{\sin^2\omega T_0/2}{(\omega T_0/2)^2}\end{aligned}$$

The spectral density is illustrated in figure 1.7.

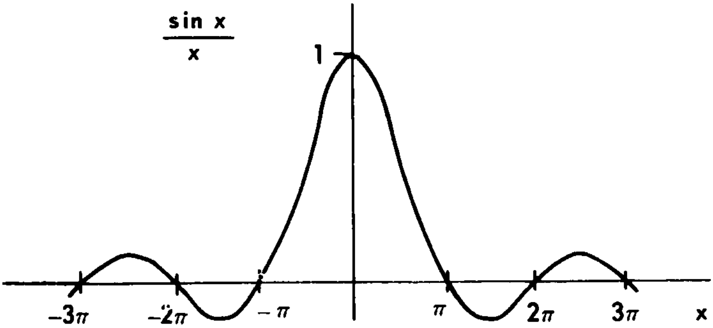


FIGURE 1.7—The function $(\sin^2 x)/x^2$.

- The spectral density decreases rapidly with ω. In fact, about 90 percent of the power is included within the frequency interval $|\omega| \leq 2\pi/T_0$. Clearly, then, passing this signal through a filter with the transfer function

$$H(j\omega) = \begin{cases} 1 & |\omega| \leq \dfrac{2\pi k}{T_0} \\ 0 & \text{Otherwise} \end{cases}$$

would alter it only slightly for k on the order of 3 or 4.

BANDWIDTH

The *bandwidth* of a signal is a measure of the width of the spectrum of the signal in cycles per second (ω rad/sec $= \omega/2\pi$ cps $= f$ cps). The bandwidth of the signal just discussed is infinite, since the power spectral density goes to zero only asymptotically with ω. This is true with most common signals. It is, therefore, useful to define an *effective* bandwidth B_{eff} such that, were the spectrum constant with the value $\Phi_y(0)$ out to some point $\omega = 2\pi B_{\text{eff}}$, the average power would be the same as in the actual signal. That is

$$P_{\text{ave}} = \int_{-2\pi B_{\text{eff}}}^{2\pi B_{\text{eff}}} \Phi_y(0) \frac{d\omega}{2\pi} = 2B_{\text{eff}}\Phi_y(0) = \int_{-\infty}^{\infty} \Phi_y(\omega)\, \frac{d\omega}{2\pi} \tag{1.50}$$

For the signal just discussed, it follows that $P_{\text{ave}} = A^2 = 2B_{\text{eff}}A^2T_0$ and hence that $B_{\text{eff}} = 1/2T_0$. Thus, the more pulses that occur each second, the greater is the signal bandwidth.

This inverse relationship between how rapidly a signal can change and its effective bandwidth can be illustrated intuitively as follows: The rapidity with which a signal changes is measured by its autocorrelation function. If $\phi_y(\tau) \approx \phi_y(0)$ then the signal looks much the same at time $t+\tau$ as it did at time t. If on the other hand $\phi_y(\tau) \approx 0$ then $y(t)$ and $y(t+\tau)$ are quite different. (We are assuming here that the average value of $y(t)$ is zero.) Thus, the smaller the value of τ necessary before $\phi_y(\tau) \approx 0$, the more rapidly the signal is changing. As a measure of this rapidity of change let us define the *time width* τ_0 analogously with the definition of effective bandwidth so that

$$\tau_0\phi_y(0) = \int_{-\infty}^{\infty} \phi_y(\tau)\, d\tau \tag{1.51}$$

But, observe that

$$\tau_0\phi_y(0)=\int_{-\infty}^{\infty}\phi_y(\tau)\,\mathrm{d}\tau=\Phi_y(0)=\frac{1}{2B_{\mathrm{eff}}}\int_{-\infty}^{\infty}\Phi_y(\omega)\,\frac{\mathrm{d}\omega}{2\pi}$$

$$=\frac{1}{2B_{\mathrm{eff}}}\,\phi_y(0)$$

and hence that

$$B_{\mathrm{eff}}=\frac{1}{2\tau_0} \tag{1.52}$$

In particular, in the case of an "almost" white process, it was observed that $\phi_y(0)=N_0W$ and hence that

$$\tau_0=\frac{1}{N_0W}\int_{-\infty}^{\infty}\phi_y(\tau)\,\mathrm{d}\tau=\frac{1}{2W}$$

Clearly, the effective bandwidth is equal to the actual bandwidth for a process with a flat spectral density.

EXPECTATION AND INDEPENDENCE

Another useful concept in connection with a random process is its time average. The time average $<y(t)>$ of a process $y(t)$ is defined as

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}y(t)\,\mathrm{d}t \tag{1.53}$$

Thus

$$<y(t)y(t+\tau)>=\phi_y(\tau)$$

and

$$<y^2(t)>=P_{\mathrm{ave}}$$

The *cross-correlation* $\phi_{y_1y_2}(\tau)$ between two functions $y_1(t)$ and $y_2(t)$ can be conveniently defined using the notation for time average

$$\phi_{y_1y_2}(\tau)=<y_1(t)y_2(t+\tau)> \tag{1.54}$$

In general, the random functions with which we shall be concerned will have zero time average

$$<y(t)>=0$$

If not, then $<y(t)>=C$ for some constant C, and we can define a new random process $y'(t) = y(t) - C$ which does have a zero average. Two random functions $y_1(t)$ and $y_2(t)$ are said to be *linearly independent* if

$$<y_1'(t)y_2'(t)> = <[y_1(t) - C_1][y_2(t) - C_2]> = 0$$

where $C_1 = <y_1(t)>$ and $C_2 = <y_2(t)>$.

Now, consider the quantity $<y(t)f(t)>$ where $y(t)$ is a stationary random process and $f(t)$ is a periodic function with period T_0. Then, formally,

$$\begin{aligned}
<y(t)f(t)> &= \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} y(t)f(t)\,dt \\
&= \lim_{N\to\infty} \frac{1}{(2N+1)T_0}\int_0^{T_0} f(t)\sum_{i=-N}^{N} y(t+iT_0)\,dt \\
&= \lim_{N\to\infty} \frac{1}{(2N+1)T_0}\sum_{j=0}^{T_0/\Delta t-1}\int_{j\Delta t}^{(j+1)\Delta t} f(t)\sum_{i=-N}^{N} y(t+iT_0)\,dt \\
&= \lim_{\Delta t\to 0}\lim_{N\to\infty} \frac{1}{(2N+1)T_0}\sum_{j=0}^{T_0/\Delta t-1} f(j\Delta t)\sum_{i=-N}^{N}\int_{j\Delta t}^{(j+1)\Delta t} y(t+iT_0)\,dt
\end{aligned}$$

But

$$\lim_{N\to\infty}\frac{1}{(2N+1)\Delta t}\sum_{i=-N}^{N}\int_{j\Delta t}^{(j+1)\Delta t} y(u+iT_0)\,du = <y(t)>$$

since $y(t)$ is stationary and the infinite sum represents an average over an infinite time interval of the function $y(t)$. Therefore,

$$\begin{aligned}
<y(t)f(t)> &= <y(t)> \lim_{\Delta t\to 0}\frac{1}{T_0}\sum_{j=0}^{T_0/\Delta t-1} f(j\Delta t)\Delta t \\
&= <y(t)>\frac{1}{T_0}\int_0^{T_0} f(t)\,dt \qquad (1.55)
\end{aligned}$$

when $y(t)$ is a stationary random process and $f(t)$ is periodic. When $f(t)$ is periodic with period T_0, $g(t) \equiv f(t)f(t+\tau) = f(t+T_0)f(t+\tau+T_0) = g(t+T_0)$ and $g(t)$ is also periodic with period T_0. Thus it follows that if $f(t)$ is periodic with period T_0

$$\begin{aligned}
&<f(t)f(t+\tau)y(t)y(t+\tau)> \\
&\qquad = <g(t)x(t)> = <x(t)>\frac{1}{T_0}\int_0^{T_0} f(t)f(t+\tau)\,dt \qquad (1.56)
\end{aligned}$$

where $y(t)$ and hence $x(t) = y(t)y(t+\tau)$ represent stationary random processes.

Another useful concept is that of *expectation*. Let $x(t)$ represent a stationary random process with the probability density function $p(x_t)$. Then $E[f(x_t)]$, the expectation of $f(x_t)$, is defined as

$$E[f(x_t)] = \int_{-\infty}^{\infty} f(x_t)p(x_t)\ dx_t \tag{1.57}$$

If, for example $f(x) = x$, we have

$$E(x_t) = \int_{-\infty}^{\infty} x_t p(x_t)\ dx_t$$

Note the similarity between this and the definition of $<x(t)>$

$$<x(t)> = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} x(t)\ dt$$

To illustrate this similarity, suppose $x(t)$ can assume only the values A and B. Then

$$\begin{aligned}<x(t)> &= \lim_{T\to\infty}\frac{1}{2T}\Big\{A\cdot\text{amount of time in } 2T \text{ seconds that } x(t) = A \\ &\qquad + B\cdot\text{amount of time in } 2T \text{ seconds that } x(f) = B\Big\} \\ &= A\cdot(\text{percentage of time } x(t) = A) \\ &\qquad + B\cdot(\text{percentage of time } x(t) = B)\end{aligned}$$

But

$$\begin{aligned}E(x_t) &= \int_{-\infty}^{\infty} x_t p(x_t)\ dx_t = A\int_{A-\epsilon}^{A+\epsilon} p(x_t)\ dx_t + B\int_{B-\epsilon}^{B+\epsilon} p(x_t)\ dx_t \\ &= A \text{ Prob } (x_t = A) + B \text{ Prob } (x_t = B)\end{aligned}$$

If $x(t)$ assumes the values A with probability p, then, on the average it will presumably be equal to A for p percent of the time. Indeed, the fraction of time a function assumes a certain value can be used as the definition of the probability of that value. Hence, intuitively

$$E(x_t) = <x(t)> \tag{1.58}$$

and, more generally

$$E[f(x_t)] = <f[x(t)]> \tag{1.59}$$

While these relationships are not true for all random processes (e.g., they certainly do not hold for nonstationary processes since $E(x_t)$ is a function of time while $<x(t)>$ is not), they are true for the class of random processes which we shall have occasion to consider here. Thus we shall use the operators $< \ >$ and $E(\ )$ interchangeably, referring to both as "the expectation of" or "the average of." Note, in particular, that when $x(t)$ is a random process and $f(t)$ a deterministic process

$$\begin{aligned} <x(t)+f(t)> &= <x(t)> + <f(t)> \\ &= E(x_t) + <f(t)> \end{aligned} \tag{1.60}$$

and, from equation (1.55)

$$\begin{aligned} <x(t)f(t)> &= <x(t)> <f(t)> \\ &= E(x_t) <f(t)> \end{aligned} \tag{1.61}$$

NOISE

Were it not for noise, space communication would be relatively simple. The receiver antenna would be followed by an amplifier or amplifiers with enough gain to render the signal useful. While the receiver might be somewhat costly for very small signals, no signal would be too small and no transmitter would be too weak or too far removed for effective communication. Unfortunately, unpredictable random phenomena other than the signal are always present in any receiver. This noise is amplified by the same factor as the signal, and, while the voltage level may be amplified to more practical ranges, the signal is just as noisy, relatively, as it was before amplification. It is therefore the *signal-to-noise ratio* that is crucial and not the signal or noise amplitude alone. Noise must be counteracted by means other than amplification. In order to understand methods by which the effect of noise can be diminished, it is necessary to consider for a moment the properties of the noise itself.

Any electronic system generates some random voltage or current fluctuations. A metalic resistor, for example, contains electrons which drift randomly from molecule to molecule. When this resistor is connected into a circuit, the electron drift will produce a random current through the resistor and hence a voltage across its terminals. Such noise is called *Johnson* or *thermal* noise. The voltage generated across the terminals of an element is dependent upon the load into which it is operating. It is, however, convenient to be able to have a quantitative measure of noise power produced by a resistor which is independent of the circuit of which it forms a part. For this reason, the noise power produced by a resistor will be identified here with the amount of noise it generates in a *matched load;* i.e., a load whose resistance is identical to that of the resistor itself. Therefore, let the random (matched) noise voltage be $v(t)$ volts at time t. Then the voltage autocorrelation function is

$$\phi_v(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} v(t)v(t+\tau)\, dt$$

and the power spectral density is

$$\Phi_v(\omega) = \int_{-\infty}^{\infty} \phi_v(\tau) e^{-j\omega\tau}\, d\tau$$

It has been determined both experimentally and theoretically that for thermal noise

$$\Phi_v(\omega) = \frac{kT}{2} \tag{1.62}$$

where k is the Boltzmann constant, $k = 1.38 \times 10^{-23}$ joule/°K, and T is the absolute temperature of the resistor in degrees Kelvin. While $\Phi_v(\omega)$ is not constant for all values of ω (this, as we have seen, would indicate infinite power), the spectral density is flat out to extremely high frequencies, on the order of 10^{13} cps, where quantum effects occur.

Another commonly encountered noise source in an electronics system is the so-called *shot noise.* This is noise produced in vacuum tubes. Many millions of electrons are emitted randomly from the cathode of an electron tube each second. The average emission rate determines the average current. Nonetheless, because of the discrete properties of an electron, this current is not continuous, but rather is composed of many electron pulses. Thus, the instantaneous current fluctuates about this average. Since this fluctuation is random and is not produced by any input signal variations, it acts as noise. Shot noise may be accounted for by an equivalent noise source producing a noise power spectral density of $\frac{1}{2}kT_0$ watts, where k is the Boltzmann constant and T_0 is the *effective* noise temperature of the tube. This spectral density is flat out to frequencies on the order of the reciprocal of the time necessary for an electron to pass from the cathode to the anode. Although this frequency varies from tube to tube, it is typically on the order 10^9 cps.

Solid-state devices, including transistors, also exhibit noise due to random electron fluctuations. Again this may be accounted for by including, in the circuitry of the transistor, an effective noise generator with a power spectral density of $\frac{1}{2}kT_t$, where k is again the Boltzmann constant, and T_t the effective noise temperature of the device. (Note that the effective noise temperature is not necessarily the actual temperature of the device in question.) The range of frequencies for which this spectral density remains constant is generally somewhat less than that for vacuum tubes. However, it is usually safe to assume that the spectrum is flat over the frequency range for which the device is useful.

In short, any electronic device generates noise, the spectral density of which is directly proportional to the product of its effective temperature (or to the temperature of one of its elements), and the Boltzmann constant k. Any electronic system contains many of these independent noise generators. Assuming the system is linear, all of these noise sources can be replaced theoretically by one noise generator at the input to the device which produces at the output a noise power equivalent to that that is actually generated by the combined action of the separate sources. The total (single-sided) noise spectral density theoretically needed at the input to account for the observed output is often written $N_0 = kT_{\text{eff}}$ where T_{eff} is designated the *effective noise temperature* of the system in degrees Kelvin. Since the input signal and the effective input noise are equally amplified by the system, the signal-to-noise ratio remains the same at the output as at the input. Thus, knowing the effective temperature of the system, it is only necessary to determine the input power in order to specify the output signal-to-noise ratio.

GAUSSIAN STATISTICS

While no attempt is made in this report to develop a background in probability theory, it is necessary to introduce a few rather fundamental concepts. As mentioned earlier, a random process can be partially characterized by the probability density of its amplitude. Thus, associated with a stationary random process, $y(t)$ is a density function $p(y_t)$ with the property that, at any instant of time t, the amplitude of y_t falls within the limits $a \leqq y_t \leqq b$ with the probability

$$Pr(a \leqq y_t \leqq b) = \int_a^b p(y_t)\ dy_t$$

That is, if $y(t)$ is sampled at a large number of times t_i, it will be found that about $100 \times Pr\{a \leqq y_t \leqq b\}$ percent of the samples will be bounded by a and b in amplitude. It will be observed that

$$Pr\{-\infty < y_t < \infty\} = \int_{-\infty}^{\infty} p(y_t)\ dy_t = 1$$

since y_t must have some amplitude.

Probability densities, of course, can take on an infinite number of functional forms. However, a powerful theorem, known as the Central Limit Theorem, states the following extremely useful result: Let x_i be a random variable with the density function $p(x_i)$. Consider a sum of these (independent) variables

$$z_N = \sum_{i=1}^{N} x_i$$

Subject to some quite general restrictions on the density function $p(x_i)$, the random variable z_N is distributed, asymptotically as $N \rightarrow \infty$, according to the density function

$$p(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} \qquad x \equiv \frac{z_N - \mu}{\sigma} \tag{1.63}$$

The term μ, called the *mean* of z_N, is the expected value of z_N and σ^2, designated the *variance* of z_N, is the expected value of $(z_N - \mu)^2$. A variable z_N with this density function is called a random variable of Gaussian or normal distribution. The function $p(x)$ is illustrated in figure 1.8(a) and its integral, the *cumulative Gaussian distribution* function

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{Z} \exp\left[-\frac{(z_N-\mu)^2}{2\sigma^2}\right] dz_N = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{X} \exp\left(-\frac{x^2}{2}\right) dx \equiv P(X) \tag{1.64}$$

where $X = (Z - \mu)/\sigma$, is plotted in figure 1.8(b). Note that $P(-\infty) = 0$, $P(0) = \frac{1}{2}$, and $P(\infty) = 1$. A convenient function used in later chapters is the *error function*, defined as

$$\begin{aligned} \text{erf}\,(X) &\equiv \frac{1}{\sqrt{2\pi}}\int_{-X}^{X} \exp\left(-\frac{x^2}{2}\right) dx = \frac{2}{\sqrt{2\pi}}\int_{0}^{X} \exp\left(-\frac{x^2}{2}\right) dx \\ &= 2[P(X) - P(0)] \end{aligned} \tag{1.65}$$

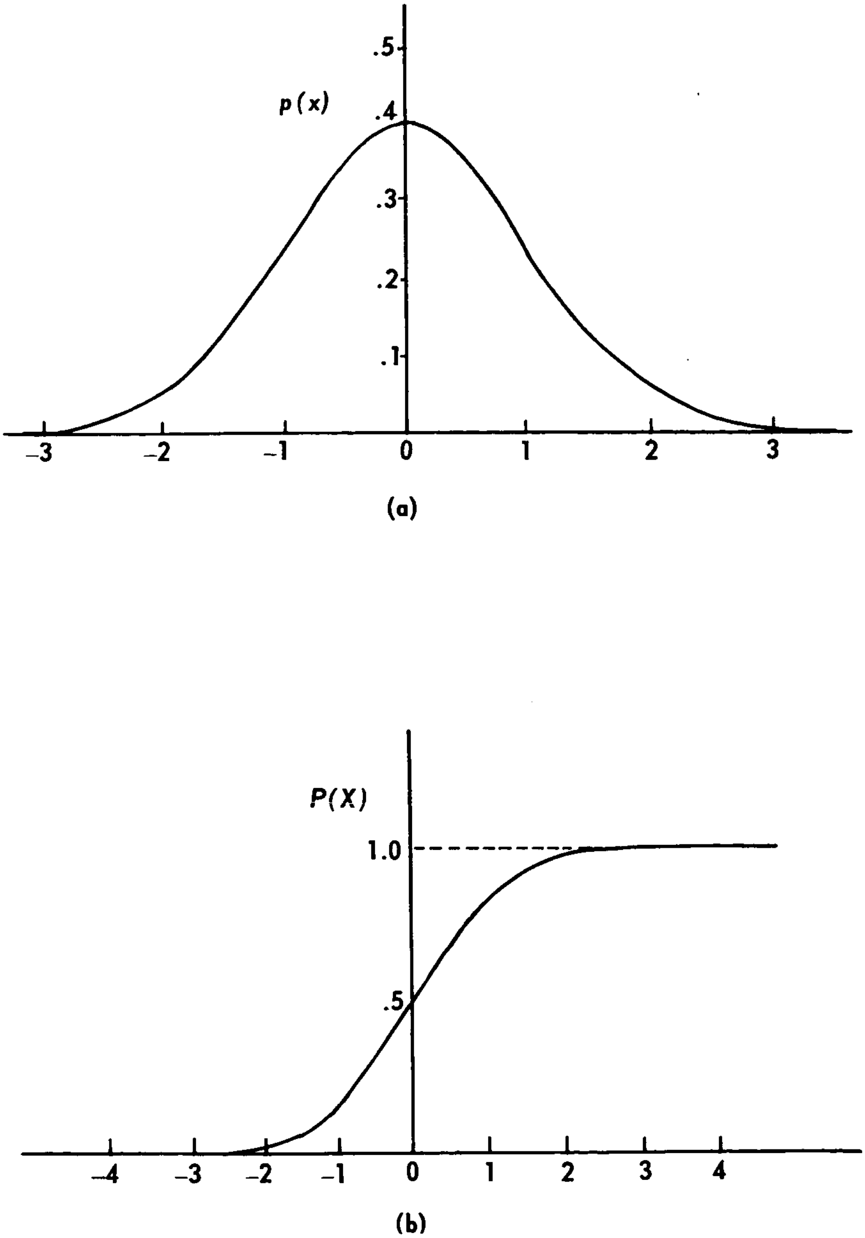
Hence

$$P(X) = \tfrac{1}{2}\,[\text{erf}\,(X) + 1]$$

and

$$\begin{aligned} Pr(x > X) &= \frac{1}{\sqrt{2\pi}}\int_{X}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx \\ &= \tfrac{1}{2}[1 - \text{erf}\,(X)] \end{aligned} \tag{1.66}$$

It was stated in the previous section that electronic noise is caused by the aggregate of a large number of random phenomena. Vast numbers of electrons are involved in producing any electronic current. While the distribution of the emission times, for example, of electrons from the cathode of a vacuum tube may not be known, one would suspect that the distribution of the amount of current produced by the electrons, since it is an effect of the combination of a large number of these random events, would be Gaussian. This is indeed true for all of the noise



(a)

(b)

FIGURE 1.8—The Gaussian probability functions. (a) The Gaussian probability density function $p(x)$; (b) the cumulative Guassian distribution function $P(X)$.

sources described in the last section. In fact, a considerably stronger statement can be made concerning these random noise phenomena: they may be characterized as *Gaussian random processes.* If a random process is Gaussian, not only is its amplitude distribution at any time instant t Gaussian, but, in addition, certain relationships between the random variables x_{t_1}, x_{t_2}, . . ., etc., must hold for all values of t_1, t_2, ... Since we cannot go into more detail here concerning random processes, the interested reader is referred to the literature (see, for example, the bibliography for this chapter). The property of a stationary Gaussian process $x(t)$ that will be of most use to us in these notes is that, in addition to x_t being a Gaussian distribution, any integral

$$z=\int_{t_a}^{t_b} a(t)x(t)\ \mathrm{d}t \tag{1.67}$$

of the process $x(t)$, where $a(t)$ is a deterministic function of time, has also a Gaussian distribution with the mean

$$\mu_t=E(z)=\int_{t_a}^{t_b} a(t)E(x_t)\ \mathrm{d}t=E(x_t)\int_{t_a}^{t_b} a(t)\ \mathrm{d}t \tag{1.68}$$

and variance

$$\begin{aligned}\sigma_t^2&=E[(z-\mu_z)^2]\\&=E(z^2)-2\mu_zE(z)+\mu_z^2\\&=E(z^2)-\mu_z^2\\&=\int_{t_a}^{t_b}\int_{t_a}^{t_b} a(t)a(u)E[x(t)x(u)]\ \mathrm{d}t\ \mathrm{d}u-\mu_z^2\end{aligned} \tag{1.69}$$

(Note from the definition of the operator E that $E(ax)=aE(x)$ where a is a constant, and $E(x+y)=E(x)+E(y)$.)

The noise produces a random variation about the signal portion of the total output. The average value of the noise itself is usually zero. The variance of the noise is the average value of the noise squared (when its mean is zero). Thus

$$\sigma^2=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} n^2(t)\ \mathrm{d}t=\phi_n(0)=<n^2(t)>=E(n_t^2) \tag{1.70}$$

Note, too, that since the noise we have described is effectively white

$$
\begin{aligned}
<n(t)n(u)> &= <n(t)n[t+(u-t)]> = \phi_n(u-t) \\
&= \int_{-\infty}^{\infty} \Phi_n(\omega)\, e^{j\omega(u-t)} \frac{d\omega}{2\pi} = \frac{N_0}{2}\int_{-\infty}^{\infty} e^{j\omega(u-t)} \frac{d\omega}{2\pi} = \frac{N_0}{2}\delta(u-t) \quad (1.71)
\end{aligned}
$$

These results prove to be quite useful in subsequent chapters.

CHAPTER 2

Amplifiers and Antennas

AS WE OBSERVED in the previous chapter, noise is an unavoidable part of any communication system. In space telemetry systems this noise is essentially additive, white, and Gaussianly distributed. Since, as we argued earlier, it is the ratio of the signal to the noise power that determines the performance of any communication system, the same result may be accomplished by either increasing the signal power at the receiver, or decreasing the noise power. In this chapter ways for accomplishing both of these tasks are discussed.

The greatest noise contribution in most space telemetry situations arises in the initial stages of the receiver. Since the transmitter operates at relatively high signal levels, the signal-to-noise ratio at the transmitter can be kept very large. Background radiation at the frequencies generally used is relatively insignificant (and, in any event, unavoidable). At the receiver, however, the signal power is extremely low so that any noise contributed in the initial process of amplifying this signal may be, in comparison, most significant. It is at this stage that the greatest effort is demanded to decrease the additive noise, and it is here that the most spectacular progress has been made.

LOW-NOISE AMPLIFIERS

Signal amplification is most commonly achieved with vacuum tube and transistor amplifiers, but because such amplifiers are relatively noisy and, at any rate, not practical at the frequency ranges used for space communication, we shall not be concerned with them here. The earliest technique for the low-noise amplification of microwave frequencies involved the use of the *traveling-wave tube.* The traveling-wave tube developed during World War II relies upon the interaction between an electron beam and the signal-bearing electromagnetic wave. This electromagnetic wave is effectively slowed down to the velocity of the electron beam by passing it through a waveguide, generally in the shape of a helix. Since electromagnetic energy traverses linearly along a waveguide at nearly the velocity of light c, its rate of progress along the axis of the helix is approximately $(l/L)c$, where l is the length of the axis of the helix, and L is the length of the waveguide comprising the helix. Thus, it is possible to make the velocities of linear propagation of the signal and the electron beam equal. When this is done, there is an interaction between the electric field of the signal and the electrons in the beam. The electrons' densities are increased or decreased, depending upon the intensity and direction of the field. This "bunching," in turn, causes the field to be intensified in proportion to its original strength, thus producing amplification. Extremely large amplifications over a wide band of microwave frequencies are, indeed, possible with this technique. The noise arises, as usual, because the electrons do not all have the same energy or velocity. Thus the bunching cannot be perfect. Since the electrons are not all moving with the same velocity, they exhibit a countertendency toward a random distribution. This appears as noise at the output. Much effort has been made to decrease the noise inherent in traveling-wave tubes, and amplifiers using these tubes have been built with effective noise temperatures of less than 300° K.

Another more recent development in low-noise, broadband microwave amplifiers is the *parametric amplifier.* The action of this device is commonly compared to the method by which a child, sitting in a swing, is able to increase the amplitude of his swinging arc. At the height of his displacement, when the swing changes directions, the child pulls back on the ropes, thereby slightly increasing his height, and hence his potential energy. At the bottom of the arc, the tension on the ropes is relaxed so that this potential energy is entirely converted into kinetic energy. Because the maximum height was increased, this kinetic energy is greater than it would have been, and at the next peak the potential energy has increased over its value at the previous peak. The energy of the child, therefore, is converted into oscillation energy of the swing.

In the same way any oscillator or resonant device can gain energy by being "pumped" at the right times. In fact, it can be shown that the oscillator exhibits a net energy gain even if it is pumped at the "wrong" time; that is, even if the pumping frequency and the swinging frequency are not in a one-to-one relationship.

This, then, is the principle of the parametric amplifier. An oscillator or resonator generally possesses two means of energy storage. If the storage capacity of one of these devices (or parameters) is altered (as the child changes the effective length of the ropes of the swing) at a frequency high compared with its natural frequency, the resonator will exhibit a net increase of energy. This increase can be sizable.

One realization of the parametric amplification principle is obtained through the use of solid-state devices which have the property that their energy storage capacity varies in inverse proportion to the intensity of the applied signal. This applied, or pumping, signal is chosen to have perhaps 10 times as great a frequency as the signal to be amplified. The latter signal, when used as in input to the resonant circuit containing the pumped element, is thereby amplified. Noise in such a device arises primarily from the dissipative elements in the storage devices and the associated circuitry. These effects can be kept to a minimum, however, by cooling the amplifier, with liquid helium, for example, to a few degrees Kelvin. It is possible, in this way, to get effective noise temperatures of 100° K or less for parametric amplifiers.

Another implementation of the parametric principle is the electron-beam parametric amplifier. Here, an electron beam is used as the energy storage device and is pumped by an electric field. The performance of this device is approximately that of the more conventional parametric amplifier.

Probably the most successful low-noise amplifier yet developed, however, is the *maser* (acronym for *m*icrowave *a*mplification by *s*timulated *e*mission of *r*adiation). The electrons in the crystal lattice of any material, like all electrons, spin about some axis. The orientation of the spin axes is restricted to certain positions, and normally the vast majority of the electrons is in the lowest energy position. If the difference in energy between the lowest two energy levels is ΔE, an amount of energy ΔE is absorbed by the crystal when an electron makes a transition from the lowest to the next lowest level, and an amount of energy ΔE is radiated when the reverse transition occurs. Normally, transitions occur equally often in each direction so that the net radiated energy is zero. The frequency of this radiation, we know from Planck's equation, must be

$$f=\frac{\Delta E}{h} \tag{2.1}$$

where $h=6.6\times10^{-34}$ joule-seconds. If the crystal is radiated with energy at the frequency $\Delta E/h$, electrons are caused to make the transition to the higher level and energy is absorbed. By irradiation with energy at a higher frequency f', it is possible to excite the electrons to a still higher energy level $\Delta E'=hf'$. By the proper selection of a crystal, it is possible to achieve a situation in which electrons, excited to the level $\Delta E'$, can decay to the level ΔE, but cannot decay further, to the ground level except in the presence of external radiation at the frequency $f=\Delta E/h$. It is thus possible to create a situation in which the majority of electrons are at the next to the lowest energy level. When this is the case, a signal at the frequency f when applied to the crystal exhibits a net increase in energy due to the preponderance of electron transitions to the ground level which it triggers. Thus, energy is transferred from the higher excitation frequency to the crystal, and from the crystal to the lower signal frequency. Again, the resulting amplification can be sizable.

The noise generated in a maser amplifier can be exceedingly small. It is due to fluctuations in the radiation field in the neighborhood of the crystal. These fluctuations can be caused by thermal agitation of the electrons causing a noise spectral density $N_{th}=kT$ where T is the actual temperature (in degrees Kelvin) of the crystal. In addition, however, radiation is emitted due to spontaneous electron transitions which, according to quantum mechanics, give rise to noise with a spectral density $N_q=(1/2)hf$. For microwave frequencies the total noise $N_0=N_{th}+N_q$ can be quite small, and masers have been built with an effective noise temperature of less than 10° K. Note, however, that the N_q term is independent of temperature and hence cannot be reduced by cooling the crystal. This term, being proportional to frequency, becomes more significant at higher frequencies. At frequencies in the visible light range, for example (masers which operate at frequencies approximating those of visible light are called lasers, the *m* of microwave becoming the *l* of light), the effective noise temperature increases to about 20 000° to 30 000° K, thus seriously counteracting some of the real advantages associated with the use of lasers in space communications.

ANTENNA GAIN

Another method for mitigating the severe conditions encountered in space communications due to the vastly increased distances between the transmitter and receiver is through the use of high gain antennas. The gain of an antenna is defined as the ratio of the maximum power intensity Φ_m (the amount of power incident upon a unit area) in any direction to the average intensity Φ_{ave}. Usually, the gain G is expressed in decibels:

$$G=10\log_{10}\left(\frac{\Phi_m}{\Phi_{\text{ave}}}\right) \tag{2.2}$$

Since we are interested in communication from a point to a point rather than from a point to many points, as in commercial radio, we clearly want a somewhat different antenna design than that commonly used in the latter case. In conventional radio transmission, it is desired to radiate equally in all horizontal directions. To accomplish this, vertical or "dipole" antennas are used with heights which are, ideally, half of the wavelength of the frequency radiated. Since they radiate horizontally, with little energy being transmitted vertically, they exhibit gains which are greater than 1; in the case of an ideal dipole antenna, the gain in the equatorial plane is 2.15 dB.

For space communication, however, it is desired to radiate energy only between the one transmitter and the one receiver. (There may be more than one receiver in practice but, at deep space distances, the earth itself is effectively one point.) It is therefore necessary to be able to direct the radiated energy in a narrow beam toward the receiver. This is most effectively accomplished by focusing the energy by means of a reflector in the shape of a paraboloid. A parabola, it will be recalled, has the geometric property illustrated in figure 2.1 and hence, to the extent that the angle of incidence of a microwave beam is equal to the angle of reflection, all energy originating at the focal point and striking the antenna will be reflected in a direction parallel to the axis of the antenna A–A'.

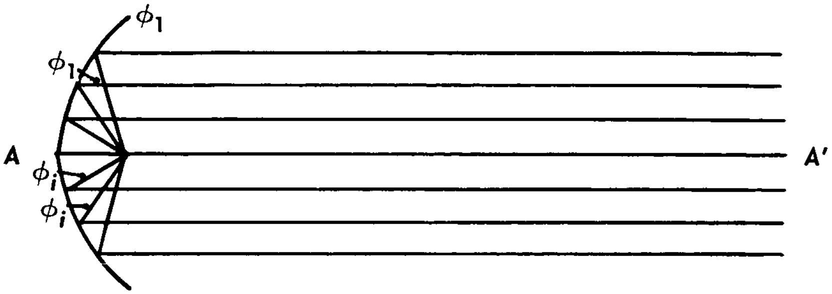


FIGURE 2.1—Direction of reflection from a parabolic antenna.

Unfortunately, however, the wavefront will not remain constant with a diameter equal to that of the antenna, but will increase in area. For an intuitive understanding of the reason for this spread, consider the illustration in figure 2.2.

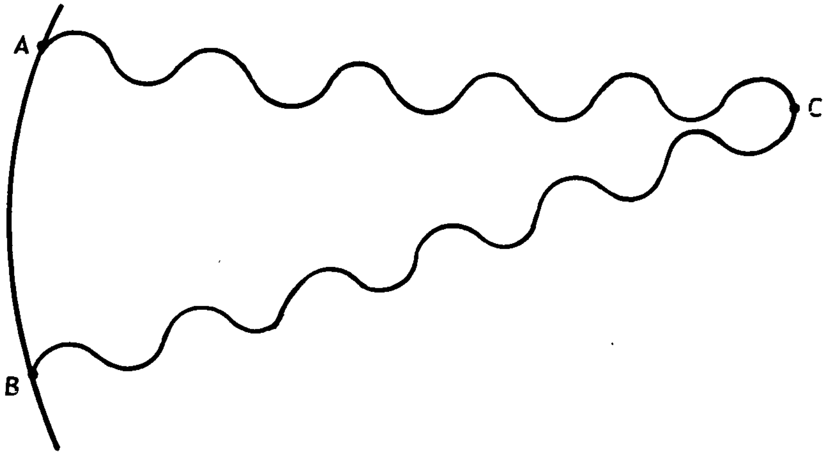


FIGURE 2.2—Interference due to reflections from different parts of the antenna.

The observer at point C "sees" energy reflected from various portions of the surface of the antenna. Consider two infinitesimal signal surface areas A and B. Since C is closer to A than to B, the energy from B must travel farther before it reaches the point C. If the geometry is such that the distance BC is exactly one-half wavelength further than the distance AC, than the radiation from the two points A and B will arrive at C exactly 180° out of phase with respect to each other. The electromagnetic fields will have equal amplitudes but opposite signs and will, therefore, completely cancel each other. When C is too close to the axis, there will be no two points on the surface of the antenna such that the difference in their distance to C is as great as one-half wavelength. As the distance from C to the axis increases, the points A and B satisfying the property described above will move closer together and, in addition, other points A' and B' can be found on the surface of the antenna such that the distances $A'C$ and $B'C$ differ by exactly three-halfs wavelength. Thus, as C moves away from the axis there is more and more cancellation, so that the net amount of energy striking C rapidly diminishes. To determine theoretically the width of the beam at a distance D from the antenna, consider the diagram in figure 2.3.

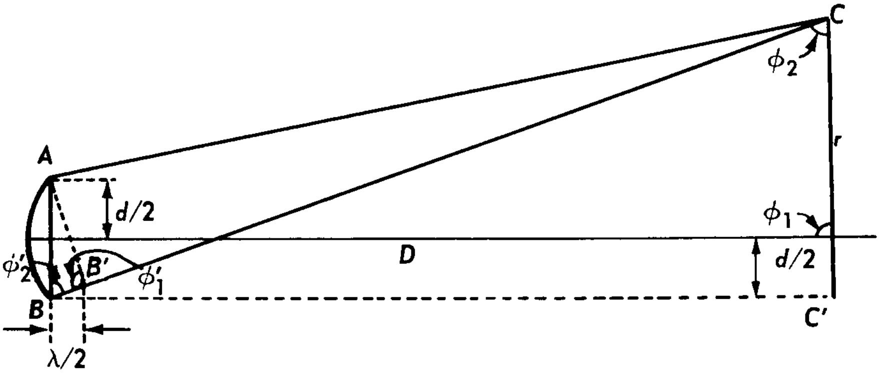


FIGURE 2.3—Determination of the beamwidth.

We want to find, as a measure of the beamwidth, the smallest value of r, the distance from C to the axis, such that there is total cancellation of energy arriving from at least two points on the antenna. Clearly, the first two points on the antenna surface which provide such cancellation are those two points in the same plane as C, and separated by the maximum distance. Therefore, let A be at one extreme of the antenna and B at the other, separated from A by a distance d, the diameter of the antenna. Assuming that D and r are large compared with the dimensions of the antenna, and that λ is small compared with d, it is easy to determine the value of r in terms of the wavelength, the antenna diameter, and the distance D. First we find the point B' on the line CB such that the distances CA and CB' are equal. Since D is large, ϕ_1' is nearly a right angle and hence $\phi_1' \doteq \phi_1$. Clearly, $\phi_2' = \phi_2$ and, hence, the triangles $BC'C$ and $AB'B$ are (nearly) similar. Thus

$$\frac{AB'}{BB'} = \frac{BC'}{CC'}$$

and since $AB' = [d^2 - (\lambda^2/4)]^{1/2} \approx d$, $BB' = \lambda/2$ (in order that we get the desired cancellation), $BC' = D$, and $CC' = r + (d/2) \approx r$, we have

$$r \approx \frac{D\lambda}{2d} \tag{2.3}$$

and the beamwidth is proportional to $D\lambda/d$.

Now consider the amount of power received by a second parabolic antenna of area A_R at a distance D from the first. Since the beamwidth is proportional to $2r$, its area is proportional to $(2r)^2$ and hence to $D^2\lambda^2/d^2$. The percentage of the power which is received, assuming $A_T < D^2\lambda^2/d^2$, is clearly proportional to the ratio of the area of the receiving surface to the area of the beam, since all the power striking the antenna surface is reflected to the focal point (the antenna is assumed to be parabolic) and hence to the receiver input. Consequently, designating by P_T the total transmitted power, and by P_R the total received power, we have

$$P_R \propto P_T \frac{A_R}{D^2\lambda^2/d^2}$$

And finally, since the area of the transmitting antenna is proportional to d^2, we have

$$P_R \propto P_T \frac{A_R A_T}{\lambda^2 D^2}$$

Actually, this heuristically derived result can be shown to be exact, so that

$$P_R = P_T \frac{A_R A_T}{\lambda^2 D^2} \tag{2.4}$$

For nonideal parabolic antennas, A_R and A_T must be replaced by an effective area which is always somewhat less than the true area. This is primarily because of the fact that it is difficult to radiate the entire surface of the antenna with energy of equal magnitude and equal phase. Typically, the effective area is 50 to 80 percent of the actual area.

As shown, the beam area at a distance D from the antenna is proportional to $D^2\lambda^2/A_T$. The power intensity over the beam front is therefore proportional to $P_TA_T/D^2\lambda^2$. If the power were radiated uniformly in all directions, the power intensity at a distance D from the antenna would be equal to $P_T/4\pi D^2$, since the power is uniformly distributed over a spherical surface of area $4\pi D^2$. Recalling the definition of gain, we have

$$G_T = 10 \log_{10}\left(\frac{\Phi_m}{\Phi_{\text{ave}}}\right) \approx 10 \log_{10}\left(\frac{4\pi A_T}{\lambda^2}\right) \tag{2.5}$$

Again, this equation can be shown to be exact so long as A_T is interpreted as the effective area. Similarly, the gain of the receiving parabolic antenna is

$$G_R = 10 \log_{10}\left[\frac{4\pi A_R}{\lambda^2}\right] \tag{2.6}$$

Then the received power in decibels is

$$10 \log_{10} P_R = 10 \log_{10} P_T + G_R + G_T - 20 \log_{10}\frac{4\pi D}{\lambda} \tag{2.7}$$

Equation (2.7) is valid regardless of the type of antennas used so long as G_R and G_T are the appropriate antenna gains.

In order to maximize the amount of power received or, equivalently, to maximize the gains of the two antennas, it is necessary to make the parabolic antennas as large as possible and the wavelengths as short as possible. First of all, there are practical limitations to the shortness of the wavelength. One of these limitations stems from the fact that, beyond a certain point, the effective noise temperature of the best amplifiers increases sharply as the wavelength decreases, thereby counteracting the advantages in antenna gain. In addition, in order to realize the theoretical gains of parabolic antennas, the dimensions of the antenna must be accurate to within a fraction of the wavelength. Since the gain increases in proportion to the area, it is advantageous to make the area as large as possible. But the larger the area, the more difficult it is to keep the tolerances within the necessary limitations. Thus, there is a tradeoff between the area and the wavelength. Moreover, because the transmitting and receiving antennas are moving with respect to each other, it must be possible to move the ground-based antenna so that this also places restrictions on its size (the fact that the vehicle antenna must be propelled through space, of course, limits its size). Finally, the transmitter antenna must be pointed in space with an accuracy proportional to the width of the beam or the maximum energy is not received at the receiver antenna. This clearly also limits the gain, and becomes particularly significant at very short wavelengths.

Antenna designs other than parabolic are also sometimes used in space telemetry. An omnidirectional antenna which, ideally, radiates or receives energy equally in all directions is always included on a spacecraft as a safety factor to enable transmission to and from the vehicle regardless of its orientation in space. Evidently this antenna has unity gain (0 dB) in all directions.

Clearly, stationary antennas can be much larger than those which must be moved. Because space telemetry antennas must be accurately pointed in space, stationary antennas are not too useful for this purpose. It is possible to get some effective direction change in stationary antennas by properly controlling the position of the source which radiates the antenna as well as the relative phases of the energy striking the various parts of the antenna surface. Such antennas are particularly useful in radio astronomy.

The gains attainable with steerable antennas can also be increased without exceeding acceptable mechanical tolerances by combining the inputs from several separate antennas. In order to realize these gains, however, the separate inputs must be accurately adjusted in phase to add constructively rather than destructively. This adjustment, of course, is a function of relative positions of the antennas and must be charged as the antennas are moved.

CHAPTER 3

Analog Modulation

BEFORE CONCENTRATING ON SOME of the recent modulation techniques for space communications, it is well to review the more conventional methods of wireless long-distance communications. Typically, a signal of the form $\sqrt{2}B \sin(\omega t+\phi)$ is generated at the transmitter. If the frequency $f=\omega/2\pi$ is sufficiently high, this signal can be applied to an antenna and will cause an electromagnetic wave to be emitted into space. A signal $\sqrt{2}A \sin[\omega(t-\tau)+\phi]$ will then be excited at the receiver antenna, where $k=A/B$ represents the attenuation due to the medium and the distance through which the signal traveled, and τ is the delay representing the time needed for the signal to travel from the transmitter to the receiver. (The medium here and throughout this report is assumed to be constant and to involve only one transmission path. In particular, k does not vary with time, except perhaps for a slow steady change due to a change in the distance between the transmitter and the receiver.)

If B and ω and ϕ are kept constant at the transmitter, virtually no information can be transmitted. The receiver is able to determine that the transmitter must exist, but essentially nothing else. If, on the other hand, any one or a combination of these parameters is varied in accordance with some rule known at both the transmitter and receiver, information can be transmitted. Commonly, there is some time function $f(t)$ which is to be transmitted representing, for example, a temperature or pressure reading on a space vehicle, or a sound or light intensity in commercial broadcasting. Thus if A in particular is made to vary proportionately with $f(t)$, $A(t)=af(t)$, the resulting *amplitude-modulated* signal is capable of conveying information. Similarly, if $\phi(t)=bf(t)$ or if $d\phi/dt=cf(t)$, the signal is said to be *phase modulated* and *frequency modulated,* respectively. These three types of modulation will be examined in some detail in the next few sections.

AMPLITUDE MODULATION

The generation of an amplitude-modulated signal is relatively straightforward. The signal $f(t)$ is converted to a voltage intensity in accordance with its amplitude (e.g., a sound wave is passed through a microphone). This voltage is amplified and multiplied by a signal $\sqrt{2}\sin(\omega_0 t+\phi)$. In addition, for reasons which will become apparent shortly, some unmodulated signal is also added. All of these procedures can be readily accomplished electronically. The resulting amplitude-modulated signal $x(t)=\sqrt{2}B[1+mf(t)]\sin(\omega_0 t+\phi)$ is then transmitted. The parameter m is referred to as the *modulation index.*

The spectrum of the signal $x(t)$ can easily be determined from the spectrum of $f(t)$. From "Spectra and Autocorrelation" in chapter 1, it will be recalled that

$$\phi_f(\tau)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}f(t)f(t+\tau)\,\mathrm{d}t \tag{3.1}$$

and

$$\Phi_f(\omega)=\int_{-\infty}^{\infty}\phi_f(\tau)\mathrm{e}^{-j\omega\tau}\,\mathrm{d}\tau \tag{3.2}$$

Thus

$$\begin{aligned}\phi_x(\tau)&=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}2B^2[1+mf(t)][1+mf(t+\tau)]\sin(\omega_0 t+\phi)\sin(\omega_0 t+\omega_0\tau+\phi)\,\mathrm{d}t\\&=2B^2<\sin(\omega_0 t+\phi)\sin(\omega_0 t+\omega_0\tau+\phi)>\\&\quad+2B^2m<f(t)\sin(\omega_0 t+\phi)\sin(\omega_0 t+\omega_0\tau+\phi)>\\&\quad+2B^2m<f(t+\tau)\sin(\omega_0 t+\phi)\sin(\omega_0 t+\omega_0\tau+\phi)>\\&\quad+2B^2m^2<f(t)f(t+\tau)\sin(\omega_0 t+\phi)\sin(\omega_0 t+\omega_0\tau+\phi)>\end{aligned}$$

It will be assumed that the expected value of $f(t)$ is zero; if not, let $f(t)=\mu+f'(t)$ where $<f(t)>=\mu$ and, hence, $x(t)=\sqrt{2}B[1+m\mu+mf'(t)]\sin(\omega_0 t+\phi)=\sqrt{2}B'[1+m'f'(t)]\sin(\omega_0 t+\phi)$. In addition, it will be recalled from "Expectation and Independence" in chapter 1 that

$$\begin{aligned}&<f(t)\sin(\omega_0 t+\phi)\sin(\omega_0 t+\omega_0 t+\phi)>\\&\qquad=<f(t)><\sin(\omega_0 t+\phi)\sin(\omega_0 t+\omega_0\tau+\phi)>=0\end{aligned}$$

and

$$\begin{aligned}&<f(t)f(t+\tau)\sin(\omega_0 t+\phi)\sin(\omega_0 t+\omega_0\tau+\phi)>=<f(t)f(t+\tau)>\\&\qquad<\sin(\omega_0 t+\phi)\sin(\omega_0 t+\omega_0\tau+\phi)>=\phi_f(\tau)\frac{\cos\omega_0\tau}{2}\end{aligned}$$

Consequently,

$$\phi_x(\tau)=B^2(\cos\omega_0\tau)[1+m^2\phi_f(\tau)]$$

and

$$\begin{aligned}\Phi_x(\omega)&=B^2\int_{-\infty}^{\infty}\left[\frac{\mathrm{e}^{-j(\omega-\omega_0)\tau}}{2}+\frac{\mathrm{e}^{-j(\omega+\omega_0)\tau}}{2}\right]\mathrm{d}\tau\\&\quad+B^2m^2\int_{-\infty}^{\infty}\left[\phi_f(\tau)\frac{\mathrm{e}^{-j(\omega-\omega_0)\tau}}{2}+\phi_f(\tau)\frac{\mathrm{e}^{-j(\omega-\omega_0)\tau}}{2}\right]\mathrm{d}\tau\\&=B^2\left\{\pi\delta(\omega-\omega_0)+\pi\delta(\omega+\omega_0)+\frac{m^2}{2}\Phi_f(\omega-\omega_0)+\frac{m^2}{2}\Phi_f(\omega+\omega_0)\right\}\end{aligned} \tag{3.3}$$

Amplitude modulate thus causes a shift in the signal spectrum with no alteration in shape. This is graphically illustrated in figure 3.1.

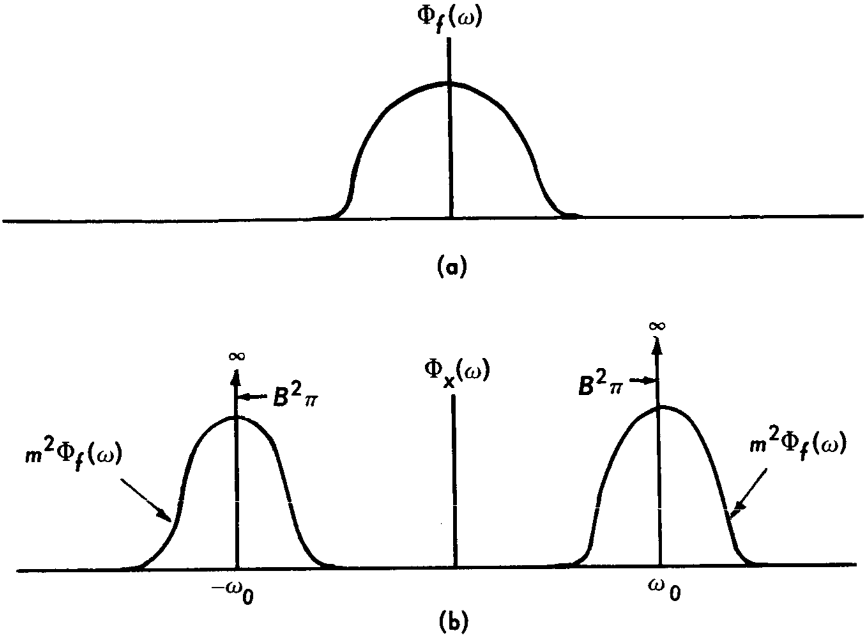


FIGURE 3.1—Spectrum shift due to amplitude modulation. (a) Modulating spectrum; (b) $\Phi_x(\omega):x(t) = \sqrt{2}B[1+mf(t)]\sin(\omega_0 t+\phi)$.

(Delta functions are conventionally represented by vertical arrows as shown, with the infinity signs designating their actual amplitude and their amplitude as drawn indicating the area under their integrals.) The average power transmitted is

$$P_{\text{ave}}=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi_x(\omega)\,d\omega=\frac{B^2}{2}+\frac{B^2m^2}{2}P_f \tag{3.4}$$

where P_f is the average power in the modulating signal $f(t)$. The percentage of power in the modulation is

$$\text{Percent power in modulation}=\frac{B^2m^2P_f}{B^2+B^2m^2P_f}=\frac{m^2P_f}{1+m^2P_f} \tag{3.5}$$

A perhaps more intuitive feeling for the spectrum of the modulated signal may be obtained by approximating the spectrum $\Phi_f(\omega)$ as shown in figure 3.2. That is, the continuous spectrum of $f(t)$ is replaced by a discrete spectrum composed of delta functions with the property that the amplitude of the delta function at the frequency $f_i=i\Delta f$ is

$$\int_{2\pi[f_i-(\Delta f/2)]}^{2\pi[f_i+(\Delta f/2)]}\Phi_f(\omega)\,\frac{d\omega}{2\pi}$$

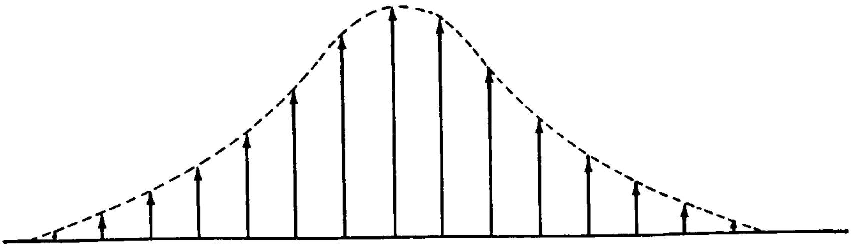


FIGURE 3.2—A discrete approximation to a continuous spectrum.

Thus, the total power remains the same, but the spectrum is assumed to contain discrete frequency components at the frequencies $f_0, \pm f_1, \pm f_2$, etc. Presumably, as $\Delta f \rightarrow 0$, the behavior predicted from the analysis of this spectrum and that corresponding to actual spectrum will be the same. Since delta functions of equal amplitude $a_i^2/4$ at the frequencies $\pm f_i$ correspond to a sinusoidal signal $f_i(t) = a_i \sin(2\pi f_i t + \phi)$, and since $\Phi_f(\omega)$ is symmetric about $\omega = 0$ (see footnote 2), $f(t)$ may be approximated by a sum of sinusoids. Consider the case when $f(t) = a \sin \omega_m t$. Then

$$\begin{aligned} x(t) &= \sqrt{2}B \sin(\omega_0 t + \phi) + \sqrt{2}Bam \sin(\omega_0 t + \phi) \sin \omega_m t \\ &= \sqrt{2}B \sin(\omega_0 t + \phi) + \frac{Bma}{\sqrt{2}} \cos[(\omega_0 - \omega_m)t + \phi] \\ &\quad - \frac{Bma}{\sqrt{2}} \cos[(\omega_0 + \omega_m)t + \phi] \end{aligned} \tag{3.6}$$

The two-sided spectrum of this signal can be written by inspection:

$$\begin{aligned} \Phi_x(\omega) &= \frac{B^2}{2} 2\pi\delta(\omega - \omega_0) + \frac{B^2}{2} 2\pi\delta(\omega + \omega_0) \\ &\quad + \frac{B^2a^2m^2}{8} 2\pi\delta(\omega - \omega_0 + \omega_m) + \frac{B^2a^2m^2}{8} 2\pi\delta(\omega + \omega_0 - \omega_m) \\ &\quad + \frac{B^2a^2m^2}{8} 2\pi\delta(\omega - \omega_0 - \omega_m) + \frac{B^2a^2m^2}{8} 2\pi\delta(\omega + \omega_0 + \omega_m) \end{aligned} \tag{3.7}$$

and is represented graphically in figure 3.3. Thus, each component of the modulating signal spectrum is translated by an amount $\omega_0/2\pi$ (and

[2] This follows from the fact that $\phi(\tau) = \phi(-\tau)$ (see "Spectra and Autocorrelation" in ch. 1). Since $\Phi(\omega) = \int_{-\infty}^{\infty} \phi(\tau)^{-j\omega\tau} \, d\tau = \int_{-\infty}^{\infty} \phi(-\tau) \, e^{-j\omega\tau} \, d\tau = \int_{-\infty}^{\infty} \phi(\tau') \, e^{j\omega\tau'} \, d\tau'$ (where we have substituted $\tau' = -\tau$), and $\Phi(-\omega) = \int_{-\infty}^{\infty} \phi(\tau) \, e^{+j\omega\tau} \, d\tau$ we have the desired result.

$-\omega_0/2\pi$) in frequency by the process of modulation. This, of course, is exactly what we concluded above.

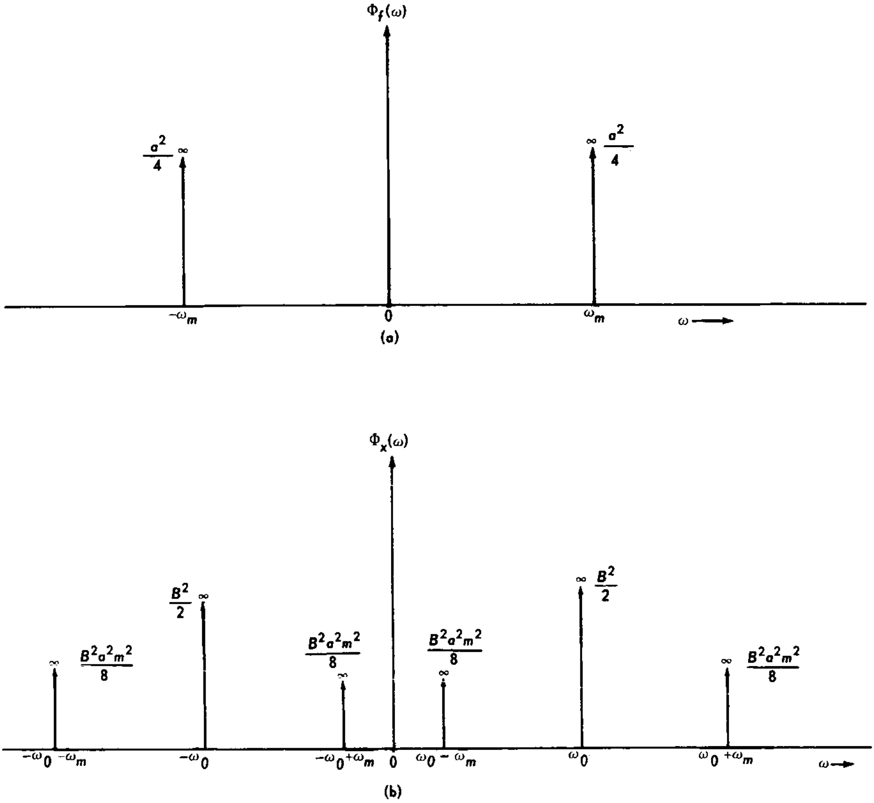


FIGURE 3.3—Two-sided signal spectrum. (a) Modulating signal spectrum; (b) modulated signal spectrum.

Suppose that the frequency $f_m = \omega_m/2\pi$ represents the highest significant frequency component in the modulating signal. Then the frequency range of the modulated signal is $f_0 - f_m$ to $f_0 + f_m$. If the bandwidth of the modulating signal is $W = f_m$ cps, then the bandwidth of the modulated signal is $[(f_0+f_m)-(f_0-f_m)] = 2W$ cps.

DEMODULATION OF AM

In order to obtain useful information from the signal $x(t)$ at the receiver, it is necessary to *demodulate* it to obtain the desired signal $f(t)$. An AM signal may be demodulated in a number of ways. The most common technique involves the use of a nonlinear element called a half-wave rectifier followed by a filter. The ideal half-wave rectifier may be regarded as a device whose output $y(t)$ is related to the input $x(t)$ as follows:

$$y(t)=\begin{cases} x(t) & x(t)\geqq 0 \\ 0 & x(t)<0 \end{cases} \tag{3.8}$$

A typical amplitude-modulated waveform is shown in figure 3.4. It is seen that so long as $mf(t)>-1$, the signal $x(t)$ is positive for $2kT<t<(2k+1)T$ and negative for $(2k+1)T<t<(2k+2)T$, for all integer values of k and for $T=2\pi/\omega_0$.

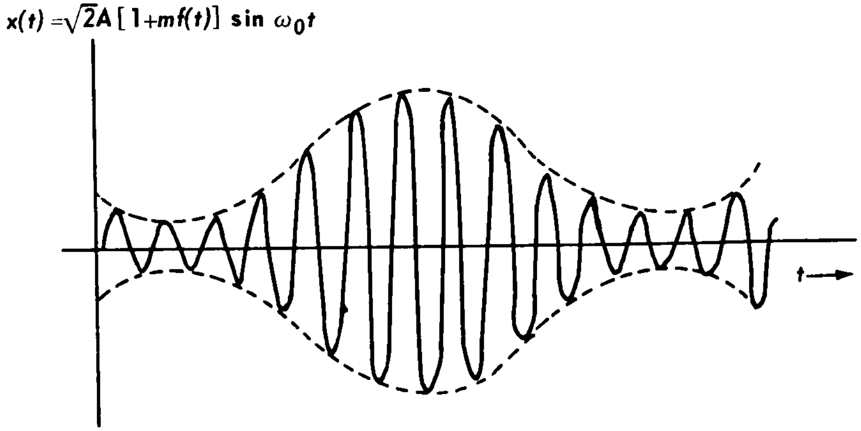


FIGURE 3.4—An AM signal.

Consequently, the half-wave rectifier has the same effect on the received waveform as if it were multiplied by the square wave shown in figure 3.5.

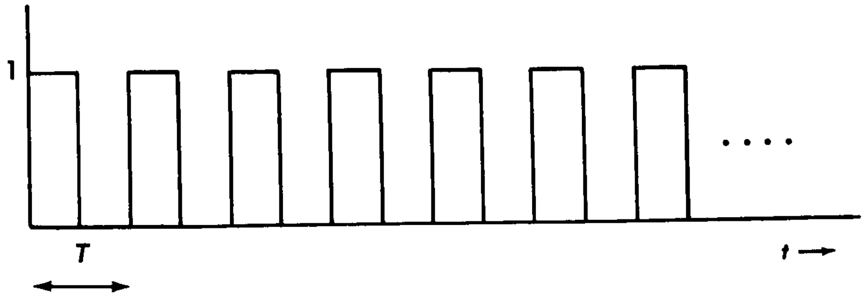


FIGURE 3.5—A square wave of period $T=2\pi/\omega_0$.

But, as shown in "Fourier Series and Fourier Transforms" in chapter 1, the Fourier series expansion of this square wave $s(t)$ is

$$s(t)=\frac{1}{2}+\sum_{n=1}^{\infty}\frac{\sin \pi n/2}{\pi n/2}\sin n\omega_0 t \tag{3.9}$$

Then

$$\begin{aligned} x(t)s(t)&=\sqrt{2}A[1+mf(t)]\sin\omega_0 t\left[\frac{1}{2}+\sum_{n=1}^{\infty}\frac{\sin \pi n/2}{\pi n/2}\sin n\omega_0 t\right] \\ &=\sqrt{2}A[1+mf(t)]\left\{\frac{1}{2}\sin\omega_0 t+\sum_{n=1}^{\infty}\frac{\sin \pi n/2}{\pi n/2}\frac{1}{2}[\cos(n-1)\omega_0 t\right. \\ &\qquad \left.-\cos(n+1)\omega_0 t]\right\} \end{aligned} \tag{3.10}$$

The product $x(t)s(t)$, therefore, contains terms centered about the frequencies 0, f_0, $2f_0$,$4f_0$, . . ., as shown in figure 3.6.

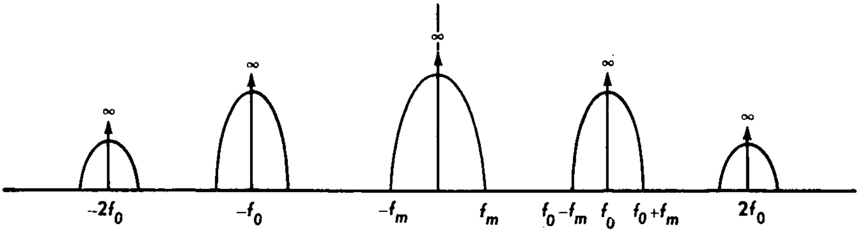


FIGURE 3.6—The spectral density of the product $x(t)s(t)$.

It is seen that all but the desired term $(\sqrt{2}A/\pi)[1+mf(t)]$ can be eliminated by filtering if $f_0-f_m>f_m$ or if $f_m<f_0/2$, where f_m is the highest frequency in the modulating signal.

DOUBLE- AND SINGLE-SIDEBAND MODULATION

While the method of AM communications described in the previous section is quite satisfactory for commercial use, it has some important limitations in those situations in which the available power is limited. First, the demodulation scheme outlined requires that $mf(t)>-1$. To appreciate the significance of this limitation, suppose $f(t)=\sin 2\pi f_m t$. Then $P_f=1/2$, and since $mf(t)>-1$, m must be less than 1, and the percentage of power in the modulation is less than 33⅓ percent (cf. eq. (3.5)).

To overcome this difficulty, consider the following demodulation scheme: A narrowband filter is centered about f_0 and the output is used to estimate the frequency and phase of the carrier. (A practical method for doing this will be considered in "Some Applications of Phase-

Locked Loops" in this chapter.) Suppose that the carrier is of the form $\sin \omega_0 t$ and the estimate $c(t) = \sqrt{2}\sin(\omega_0 t+\theta)$ is made where presumably θ is small. Then the signal can be demodulated by forming the product

$$\begin{aligned} x(t)c(t) &= 2A[1+mf(t)] \sin \omega_0 t \sin (\omega_0 t+\theta) \\ &= A[1+mf(t)][\cos \theta - \cos (2\omega_0 t+\theta)] \end{aligned}$$

which, after filtering, yields the desired signal

$$x(t)c(t) = A[1+mf(t)]\cos \theta \qquad (3.11)$$

Note that no limitations have been placed on the maximum value of $mf(t)$. The only requirement now is that there is enough power in the carrier to enable a good estimate of its frequency and phase. It is not obvious that this is superior to the previous AM system until we determine how much power must be included in the carrier for satisfactory results. In "Some Amplifications of Phase-Locked Loops" in this chapter, we shall verify, however, that in a typical situation, less than 1 percent of the total power need be included in the carrier, thus allowing a substantial increase in performance over conventional AM. This technique of increasing the proportion of power in the modulation by suppressing the carrier is commonly referred to as double-sideband, suppressed-carrier (DSB/SC) modulation.

An interesting modification of the double-sideband amplitude modulation system is afforded by a technique known as single-sideband modulation (SSB). Recalling from "Amplitude Modulation" in this chapter that the power spectrum $\Phi(\omega)$ of a real time function $f(t)$ is symmetric about $\omega = 0$, the spectrum of a typical modulating signal is as illustrated in figure 3.7(a). It has been assumed that the lowest frequency component in the modulating signal is $f_l > 0$, a situation commonly encountered in practice and one necessary for SSB modulation to be practical. As before, the upper frequency of the modulating signal is $f_m < f_0/2$. Conventional AM or DSB modulation involves the product $f(t) \sin \omega_0 t$, which, as seen in "Amplitude Modulation" in this chapter, simply shifts the spectrum to that shown in figure 3.7(b). Now suppose the signal $f(t)$ $\sin \omega_0 t$ is passed through an ideal bandpass filter with the passband $f_0+f_l<|f|<f_0+f_m$. The output then has the spectrum illustrated in figure 3.7(c). If this filtered signal $f'(t)$ is transmitted and demodulated by forming the product $f'(t) \sin \omega_0 t$ and filtering out the high-frequency components, the resulting signal has the spectrum shown in figure 3.7(d). Although two signals which have the same power spectra can be quite different, it is intuitively clear and readily verified from the manner in which the signal corresponding to the spectrum of figure 3.7(d) was formed that the particular signal is identical to the original modulating

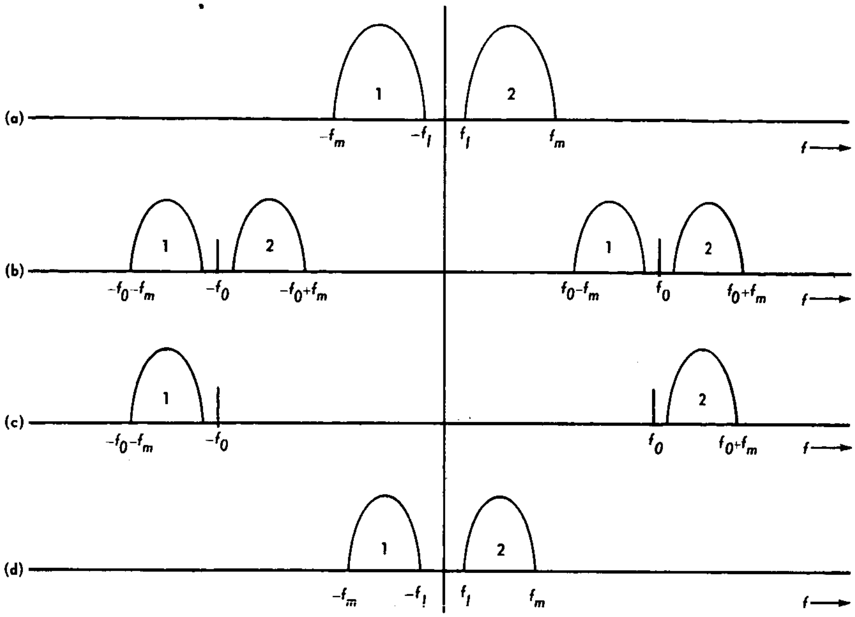


FIGURE 3.7—Power spectra illustrating the philosophy of SSB.

signal $f(t)$. But note that by filtering before transmission, as described, only half as much bandwidth is needed for SSB as for DSB modulation. As with DSB modulation, it is necessary to transmit some carrier power in order to demodulate an SSB signal. It can be shown in the latter case, however, that the phase accuracy of the estimate need not be as great as before to insure the same performance.

NOISE ANALYSIS OF AMPLITUDE MODULATION COMMUNICATION

The ultimate evaluation of any communication system rests in its behavior in the presence of noise. A convenient measure of this behavior is the output signal-power-to-noise-power ratio S/N. In the case of DSB and SSB modulation combined with product demodulation, this ratio is readily determined. Consider first DSB[3] modulation. This received signal may be written

$$x(t) = \sqrt{2}a \sin \omega_0 t + \sqrt{2}Af(t) \sin \omega_0 t$$

The total power in the modulation is A^2P_f, where P_f represents the power in the modulating signal. The signal $x(t)$ is demodulated by forming the

[3] When we refer to DSB modulation here, we intend double-sideband suppressed-carrier modulation. The DSB/SC designation is somewhat redundant. Non-suppressed-carrier modulation is denoted as "conventional AM."

product $\sqrt{2}x(t)$ sin $\omega_0 t$ and passing it through a low-pass filter with the cutoff frequency $B=f_m$. The output due to the signal is therefore

$$\{\sqrt{2}Af(t) \sin \omega_0 t \sqrt{2} \sin \omega_0 t\}_{\text{lf}} = Af(t)$$

where the subscript lf designates the low-frequency components only. The output signal power is consequently A^2P_f, the power in the modulation. The output noise signal is

$$n_1(t) = \sqrt{2}n(t) \sin \omega_0 t$$

which, as we have seen, represents a frequency translation of the noise $n(t)$. Since the input noise is white, it remains white after the product is formed:

$$\begin{aligned}&<2n(t)n(t+\tau) \sin \omega_0 t \sin \omega_0(t+\tau)>\\ &\qquad = 2<n(t)n(t+\tau)><\sin \omega_0 t \sin \omega_0(t+\tau)> = (N_0/2)\delta(\tau)\end{aligned}$$

The power spectral density of $n_1(t)$ is therefore flat with the amplitude $N_0/2$. The output noise power is consequently $(N_0/2)2B=N_0f_m$ and the output signal-to-noise ratio for DSB modulation is

$$\left(\frac{S}{N}\right)_{\text{DSB}} = \frac{A^2P_f}{N_0f_m} = \frac{P_T-a^2}{N_0f_m} \tag{3.12}$$

where P_T is the total received signal power, $P_T=A^2P_f+a^2$. When SSB modulation is used, although half the signal spectrum is suppressed, the other half can represent twice the power as before, keeping the total radiated power the same. After forming the product $\sqrt{2}x_{\text{SSB}}(t)$ sin $\omega_0 t$ and filtering, as before, it is evident from figure 3.7(d) that the situation is identical to that for DSB modulation. Hence

$$\left(\frac{S}{N}\right)_{\text{SSB}} = \frac{P_T-b^2}{N_0f_m} \tag{3.13}$$

where b is the amplitude of the received unmodulated carrier. Since, generally, a^2 and b^2 can both be small compared to the modulation power,

$$\left(\frac{S}{N}\right)_{\text{DSB}} \doteq \frac{P_T}{N_0f_m} \doteq \left(\frac{S}{N}\right)_{\text{SSB}} \tag{3.14}$$

Note that in each case we are considering ideal systems in which the unmodulated carrier power is negligible. The received signal $x(t)$ is demodulated by forming the product $\sqrt{2}x(t)$ sin $\omega_c t$. Of course, the demodulation scheme using a half-wave rectifier will not achieve the performance indicated here (although at high signal-to-noise ratios, the two methods give essentially the same results for conventional AM). Because we are considering modulation for space communications and not for commercial radio and television, we are concerned primarily with how well a particular modulation scheme can be made to work, not how well it works with inexpensive, mass-produced receivers. Thus we are only interested in the ideal system as analyzed above, which can, by the way, be approached quite closely in practice. This approach spares us the considerably greater difficulty of analyzing the signal-to-noise ratios resulting from the use of more common demodulators such as the half-wave rectifier.

ANGLE MODULATION

In this section we will consider communication systems in which the signal

$$\sqrt{2}B \sin \theta(t) \tag{3.15}$$

is transmitted, with the *angle* $\theta(t)$ varying in accordance with the modulating signal. If we define the instantaneous frequency as the rate of change of the phase angle $\theta(t)$, then $\omega(t) = \mathrm{d}\theta(t)/\mathrm{d}t$. Note that this definition corresponds to the intuitive notion of frequency when $\theta(t) = \omega t + \theta_0$. When ω varies with time, however, the intuitive definition of frequency is somewhat less clear.

A phase modulation system is one in which the phase angle $\theta(t)$ is allowed to vary with the modulating signal $f(t)$:

$$\theta(t) = \omega_0 t + \theta_0 + \Delta\theta f(t) \tag{3.16}$$

Frequency modulation, on the other hand, implies that the instantaneous frequency is made to vary with $f(t)$:

$$\omega(t) = \omega_c + \Delta\omega f(t) \tag{3.17}$$

But since $\omega(t) = \mathrm{d}\theta(t)/\mathrm{d}t$

$$\theta(t) = \int \omega(t)\,\mathrm{d}t = \omega_c t + \theta_0 + \Delta\omega \int f(t)\,\mathrm{d}t \tag{3.18}$$

then FM is essentially PM with the exception that the modulating signal in the latter is the derivative of that in the former. For this reason, the two types of modulation may be analyzed simultaneously so long as this difference is borne in mind.

As with any modulation scheme, one of the first considerations when an FM (or PM) signal is to be transmitted is that of its bandwidth occupancy. Unfortunately, the bandwidth determination for FM is considerably more difficult than that for AM. A useful simplification, when the exact shape of the spectrum is less important than its width, is to consider the case in which the modulating signal is a sinusoid

$$f(t) = \cos \omega_m t$$

Then

$$\omega(t) = \omega_c + \Delta\omega \cos \omega_m t$$

and

$$\theta(t) = \omega_c t + \frac{\Delta\omega}{\omega_m} \sin \omega_m t + \theta_0$$

and hence

$$x(t) = \sqrt{2}B \sin [\omega_c t + \beta \sin \omega_m t + \theta_0] \tag{3.19}$$

where $\beta = \Delta\omega/\omega_m$. By a trigonometric identity

$$x(t) = \sqrt{2}B \sin (\omega_c t + \theta_0) \cos [\beta \sin \omega_m t] + \sqrt{2}B \cos (\omega_c t + \theta_0) \sin [\beta \sin \omega_m t] \tag{3.20}$$

Thus, $x(t)$ may be considered to be the sum of two amplitude-modulated signals,

$$x(t) = \sqrt{2}A(t) \sin (\omega_c t + \theta_0) + \sqrt{2}B(t) \cos (\omega_c t + \theta_0)$$

To determine the spectrum of $x(t)$, we need only find the spectra of $A(t) = \cos [\beta \sin \omega_m t]$ and $B(t) = \sin [\beta \sin \omega_m t]$ and shift them by an amount ω_c as in the case of amplitude modulation.

But, since $\sin [\omega_m(t+T)] = \sin \omega_m t$ for $T = 2\pi/\omega_m$, then

$$\cos [\beta \sin \omega_m t] = \cos [\beta \sin \omega_m(t+T)]$$

and

$$\sin [\beta \sin \omega_m t] = \sin [\beta \sin \omega_m(t+T)]$$

and both terms are periodic with period T. Consequently, they may be expanded in a Fourier series; the power spectra are then determined from the squares of the Fourier coefficients. Since

$$\left.\begin{aligned} \cos [\beta \sin \omega_m t] &= \text{Re } e^{j\beta \sin \omega_m t} \\ \text{and} \quad \sin [\beta \sin \omega_m t] &= \text{Im } e^{j\beta \sin \omega_m t} \end{aligned}\right\} \tag{3.21}$$

where Re(z) and Im(z) designate, respectively, the real and imaginary parts of z, it is sufficient to expand the term $e^{j\beta \sin \omega_m t}$ in a Fourier series. We obtain

$$e^{j\beta \sin \omega_m t} = \sum_{n=-\infty}^{\infty} \frac{c_n}{T} e^{j\omega_n t} \qquad \omega_n = \frac{2\pi n}{T} \tag{3.22}$$

where

$$c_n = \int_{-T/2}^{T/2} e^{j(\beta \sin \omega_m t - \omega_n t)}\, dt \qquad \omega_m = \frac{2\pi}{T} = \omega_1$$

Letting $\zeta = 2\pi t/T$, we have

$$\frac{c_n}{T} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j(\beta \sin \zeta - n\zeta)}\, d\zeta \tag{3.23}$$

This integral commonly occurs in physical problems and has been thoroughly investigated and tabulated. It is known as the Bessel function of the first kind and is usually denoted $J_n(\beta)$, where n and β are the two parameters of c_n/T. First, we observe that, letting $\eta = \pi - \zeta$

$$J_{-n}(\beta) = \frac{1}{2\pi} \int_0^{2\pi} e^{j(\beta \sin \eta - n\eta)}\, e^{jn\pi}\, d\eta$$

$$= \frac{(-1)^n}{2\pi} \int_0^{2\pi} e^{j(\beta \sin \eta - n\eta)}\, d\eta$$

But

$$J_n(\beta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j(\beta \sin \eta - n\eta)}\, d\eta = \frac{1}{2\pi} \int_0^{2\pi} e^{j(\beta \sin \eta - n\eta)}\, d\eta$$

since $e^{j(\beta \sin \eta - n\eta)}$ is periodic in η with period 2π and all integrals over a complete period are equal, regardless of the limits. Thus

$$J_{-n}(\beta) = (-1)^n J_n(\beta) \tag{3.24}$$

and

$$e^{j\beta \sin \omega_n t} = \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j\omega_n t}$$

$$= J_0(\beta) + 2 \sum_{\substack{n=2 \\ \text{even } n}}^{\infty} J_n(\beta) \cos \omega_n t + 2j \sum_{\substack{n=1 \\ \text{odd } n}}^{\infty} J_n(\beta) \sin \omega_n t \tag{3.25}$$

As a result

$$\cos\,[\beta \sin \omega_m t] = \text{Re}\; e^{j\beta \sin \omega_m t} = J_0(\beta) + 2 \sum_{\substack{n=2 \\ \text{even } n}}^{\infty} J_n(\beta) \cos \omega_n t \tag{3.26}$$

and

$$\sin\,[\beta \sin \omega_m t] = \text{Im}\, e^{i\beta \sin \omega_m t} = 2\sum_{\substack{n=1\\ \text{odd } n}}^{\infty} J_n(\beta) \sin \omega_n t$$

and finally

$$\begin{aligned} x(t) &= \sqrt{2}B \cos\,[\beta \sin \omega_m t] \sin\,(\omega_c t+\theta_0)+\sqrt{2}B \sin\,[\beta \sin \omega_m t] \cos\,(\omega_c t+\theta_0) \\ &= \sqrt{2}BJ_0(\beta)\,(\sin \omega_c t+\theta_0)+\sqrt{2}B\sum_{\substack{n=2\\ \text{even } n}}^{\infty} J_n(\beta)\{\sin\,[(\omega_c-\omega_n)t+\theta_0] \\ &\quad + \sin\,[(\omega_c+\omega_n)t+\theta_0]\} \\ &\quad -\sqrt{2}B\sum_{\substack{n=1\\ \text{odd } n}}^{\infty} J_n(\beta)\{\sin\,[(\omega_c-\omega_n)t+\theta_0]-\sin\,[(\omega_c+\omega_n)t+\theta_0\}] \end{aligned} \tag{3.27}$$

The power at the frequency $\omega=\omega_c+\omega_n=\omega_c+n\omega_m$ is just $B^2J_n^2(\beta)$.

Since the power at a given frequency is dependent upon the magnitude of a Bessel function, it is necessary to make some observations concerning these magnitudes. In particular note that for $n>>\beta$

$$J_n(\beta) \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-in\eta}\, d\eta = 0 \tag{3.28}$$

It may be verified, by referring to a tabulation of the Bessel functions (see refs. 1 and 2), that $J_n(\beta)<0.01$ for $n>k\beta$ where $k(\beta)$ has the form shown in figure 3.8. Thus k rapidly approaches 1, and for large β the terms $J_n(\beta)$ are negligible for $n>\beta$

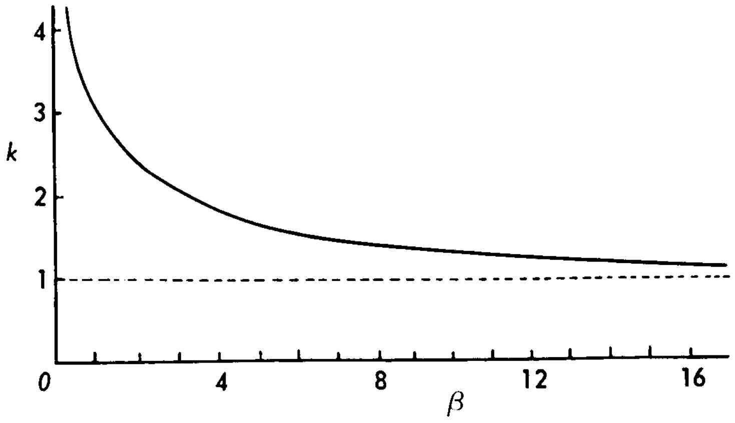


FIGURE 3.8—Values of $n/\beta=k$ such that $J_n\ (\beta)<0.01$ for $n\geq k\beta$.

The power in the received FM signal $x(t)=\sqrt{2}A \sin [\omega_c t+\beta \sin \omega_m t+\theta_0]$ is

$$\frac{1}{T}\int_0^T x^2(t)\, dt=\frac{A^2}{T}\int_0^T dt-\frac{A^2}{T}\int_0^T [\cos (2\omega_c t+2\beta \sin \omega_m t+2\theta_0)]\, dt \tag{3.29}$$

where T is the period of the function $\sin [\omega_c t+\beta \sin \omega_m t+\theta_0]$. (If ω_c is not a rational multiple of ω_m, T will be infinite.) Since the integrand of the second integral on the right of equation (3.29) is periodic with period $T/2$, the integral is zero and the total received power is A^2. Since $J_n(\beta)<0.01$ for $n\geq k\beta$, the power at the frequency component $f_c\pm nf_m$ represents less than 0.01 percent of the total power $(J_n{}^2(\beta)<10^{-4})$. Consequently, the elimination of frequencies outside the region $f_c-k\beta f_m<|f|<f_c+k\beta f_m$ should have a negligible effect upon the signal $x(t)$. The bandwidth of such an FM signal is then $W=2k\beta f_m$ and, since, for $\beta>>1, k\approx 1, W=2\beta f_m$ for large modulation indices β. That W cannot be made significantly less than this without serious distortion may be verified by again referring to a table of Bessel functions and observing that $J_n(\beta)$ increases from zero fairly rapidly as n decreases below $k\beta$.

Note that since $\beta=\Delta\omega/\omega_m=\Delta f/f_m$ in the case of frequency modulation, $W=2\Delta f$ and is independent of frequency so long as Δf, the amplitude of the modulating signal, does not vary with frequency. Thus if the average power in the modulating signal $f(t)$ is the same for all modulating frequencies (i.e., if the power spectrum of $f(t)$ is flat), then, on the average, the bandwidth occupancy of the FM signal will not vary with the frequency of the modulating signal. Since, as we shall show, the performance of FM is proportional to its bandwidth, it is desirable to have maximum bandwidth occupancy as consistently as possible.

With a phase-modulated signal, the analysis is identical except that now $\beta=\Delta\theta$ and $W=f_m\Delta\theta$. Thus, if the amplitude of the modulating signal is independent of frequency, the bandwidth of a phase-modulated signal increases with the modulating frequency, a generally less desirable situation. On the other hand, ordinary speech and music exhibit the property that the amplitude of a frequency component, beyond a certain frequency, tends to be inversely proportional to the frequency. In this case, $\Delta\theta \propto (1/f_m)$ and the bandwidth W of a PM signal remains constant, independent of frequency, whereas an FM bandwidth would decrease with increasing frequency. For this reason, commercial FM modulating signals are preceded by a preemphasis network which increases the magnitude of the higher frequency components by an amount proportional to their frequency (i.e., the modulating signal is partially differentiated). This is then counteracted by a deemphasis network at the receiver which reverses the operation. Commercial FM therefore is strictly neither FM nor PM, but a combination of both. Clearly, the distinction is irrelevant so far as the system is concerned, the only difference between the two being that of the preconditioning of the modulation signal.

Generally, then, an FM or PM modulation system is as illustrated in figure 3.9. The *voltage-controlled oscillator* (VCO) is a sinusoidal oscillator, the output frequency of which is proportional to the input voltage; if the input voltage is $f(t)$ volts, the output frequency is $f_c+f(t)\Delta f$ cps.

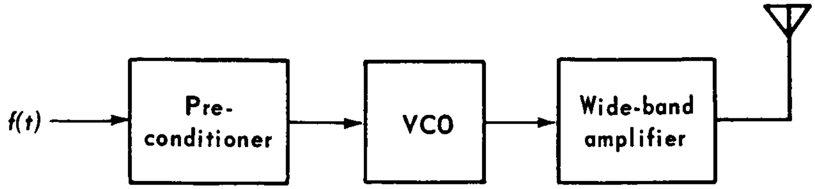


FIGURE 3.9—An FM transmitter.

There are a number of ways of implementing the block diagram of figure 3.9. However, since we are interested primarily in the system rather than in its particular realization, suffice it to observe that voltage-controlled oscillators can be designed to give the desired performance over a wide frequency range.

DEMODULATION OF ANGLE-MODULATED SIGNALS

There are several methods by which an FM signal may be demodulated. Any device capable of linearly converting a frequency variation into an amplitude variation can serve as an FM demodulator. Such a device is called a *frequency discriminator.* Suppose, for example, that the FM signal is passed through a filter with the characteristics

$$|H(j2\pi f)|=Kf \qquad f_c-\Delta f<|f|<f_c+\Delta f$$

Clearly, the output amplitude is proportional to the input frequency as desired and the FM signal is thereby demodulated. This, in fact, is a somewhat simplified version of a commercial FM discriminator.

Another FM demodulator can be designed from the following point of view: Suppose we have, at the receiver, a VCO which is identical to that at the transmitter. If we then make a preliminary estimate of the amplitude of the modulating signal and apply it to the VCO, the similarity between the output of the VCO and the received FM signal will provide us with a measure of the accuracy of this estimate. We could then use this comparison to improve our original estimate. If we use the comparison itself to adjust the VCO, the system can be made to *track* the modulation signal. One way in which this may be accomplished is illustrated in figure 3.10:

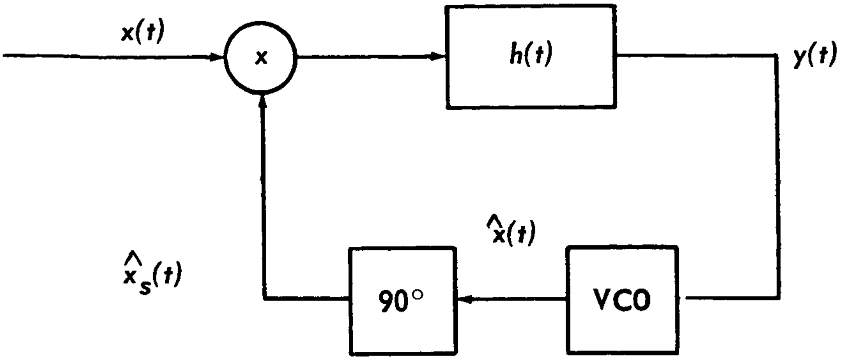


FIGURE 3.10—A phase-locked loop.

This device, called a *phase-locked* loop, consists of a multiplier, a filter $h(t)$, a VCO, and a device which shifts the phase of the VCO output by 90°. To analyze its behavior, suppose that the signal $x(t)$ is

$$x(t)=\sqrt{2}\sin(\omega_c t+\theta_1)$$

and suppose that the VCO output is

$$\hat{x}(t)=\sqrt{2}\sin(\omega_c t+\theta_2)$$

where $\theta_e=\theta_1-\theta_2$ represents a small tracking error. Then the product $x(t)\hat{x}_s(t)$, where $\hat{x}_s(t)$ represents the shifted version of $\hat{x}(t)$, is formed, yielding

$$\begin{aligned}x(t)\hat{x}_s(t)&=2\sin(\omega_c t+\theta_1)\sin\left(\omega_c t+\theta_2+\frac{\pi}{2}\right)\\&=\cos\left(\theta_e-\frac{\pi}{2}\right)-\cos\left(2\omega_c t+\theta_1+\theta_2+\frac{\pi}{2}\right)\end{aligned}$$

The last term is a high-frequency component and will be eliminated by the combined action of the VCO and the filter $h(t)$. The low-frequency component $\cos[\theta_e-(\pi/2)]=\sin\theta_e\approx\theta_e$ (the last step follows from the assumption that the phase error θ_e is small) is the input to the VCO. Suppose θ_e is positive. Then the VCO frequency is increased to something slightly greater than ω_c, thereby decreasing the difference between θ_1 and θ_2 and hence decreasing θ_e. Similarly, if θ_e is negative, the VCO frequency is decreased, again decreasing the absolute value of the difference between θ_1 and θ_2. The loop therefore acts so as to reduce the phase error to zero.

Now suppose θ_1 varies with time, $\theta_1=\theta_1(t)$. The loop will again act in such a way as to keep the phase error nearly zero. Then $\theta_2(t)\approx\theta_1(t)$. The difference between the VCO center frequency ω_c and its actual frequency is proportional to its voltage input. Since the instantaneous frequency of the VCO output is

$$\frac{d}{dt}[\omega_c t+\theta_2(t)]=\omega_c+\frac{d\theta_2(t)}{dt}$$

the input to the VCO must have amplitude

$$\frac{d\theta_2(t)}{dt}\approx k\frac{d\theta_1(t)}{dt}$$

where k is a constant of proportionality. Consequently, if the input to the loop is a frequency-modulated signal, $\theta_1(t)=\Delta\omega\int f(t)\,dt+\theta_0$, and the input to the VCO is just

$$y(t)\approx k\frac{d\theta_1(t)}{dt}=k\Delta\omega f(t) \tag{3.30}$$

and the desired signal is recovered.

ANGLE MODULATION NOISE ANALYSIS

To determine the effect of noise at the input, let $x(t)=n(t)$ be white noise. Then $n(t)\hat{x}_s(t)=n(t)\sqrt{2}\sin[\omega_c t+\theta_2(t)+(\pi/2)]=n_1(t)$ is also white noise, since as observed in "Amplitude Modulation" in this chapter, multiplying a signal by a sinusoid serves to shift its frequency spectrum. But since the spectrum of white noise is flat at all frequencies, shifting it in frequency by any amount does not alter this fact. Further, since $<n_1^2(t)> = <n^2(t)> <2\sin^2[\omega_c t+\theta_2(t)+(\pi/2)]> = <n^2(t)>$ the power spectral density retains the same magnitude before and after the multiplier (cf. "Noise Analysis of Amplitude Modulation Communication" in this chapter).

Now consider the situation in which $x(t)=\sqrt{2}A\sin[\omega_c t+\theta_1(t)]+n(t)$ and $\hat{x}_s(t)=\sqrt{2}\sin[\omega_c t+\theta_2(t)+(\pi/2)]$, where, again, it is assumed that $\theta_1(t)-\theta_2(t)$ is small. The low-frequency term of the product $x(t)\hat{x}_s(t)$ is given by

$$[x(t)\hat{x}_s(t)]=A\sin[\theta_1(t)-\theta_2(t)]+n_1(t)\approx A[\theta_1(t)-\theta_2(t)]+n_1(t) \tag{3.31}$$

Since $\theta_2(t)$ is adjusted by the action of the loop to keep the error signal and hence the input to $h(t)$ small, it follows that

$$\theta_2(t) \approx \theta_1(t) + \frac{n_1(t)}{A} \tag{3.32}$$

and, consequently

$$y(t) = k\frac{d\theta_2(t)}{dt} \approx k\frac{d\theta_1(t)}{dt} + \frac{k}{A}\frac{dn_1(t)}{dt}$$

$$= k\Delta\omega f(t) + \frac{k}{A}\frac{dn_1(t)}{dt} \tag{3.33}$$

Since the desired output is $k\Delta\omega f(t)$, the term $(k/A)(dn_1(t)/dt)$ represents output noise.

To gage the magnitude of the derivative of the noise, we approximate, as in "Amplitude Modulation" in this chapter, the continuous noise by the discrete components $a_\nu \sin(2\pi\nu\Delta ft+\theta_\nu)$ at the frequencies $f_\nu = \nu\Delta f$, $\nu = 0, 1, \ldots$. Then the derivative of the noise consists of the components

$$2\pi f_\nu a_\nu \cos(2\pi\nu\Delta ft+\theta_\nu) \tag{3.34}$$

and if the power spectral density of the noise is $\Phi(\omega) = N_0/2$, the power spectral density of its derivative is $\Phi'(\omega) = (N_0/2)(2\pi f)^2$.

The signal-to-noise ratio at the output of the FM demodulator is determined as follows: The signal power is

$$<k^2(\Delta\omega)^2 f^2(t)> = k^2(\Delta\omega)^2 P_f \tag{3.35}$$

and the noise power is

$$\frac{k^2}{A^2}\int_{-2\pi W}^{2\pi W} \Phi'(\omega)\frac{d\omega}{2\pi} = \frac{k^2}{A^2}\frac{N_0}{2}\int_{-W}^{W}(2\pi f)^2\, df$$

$$= (2\pi)^2\frac{k^2 N_0 W^3}{3A^2} \tag{3.36}$$

where W is the bandwidth of the output signal. Clearly, $W = f_M$, the maximum frequency component of the modulating signal, since no higher frequencies are of interest. (If the loop itself did not eliminate all frequencies greater than f_M cps, it could be followed by a low-pass filter which did.) Therefore the output signal-to-noise ratio is

$$\begin{aligned}\left(\frac{S}{N}\right)_{\text{FM}} &= \frac{3A^2(\Delta\omega)^2P_f}{(2\pi)^2N_0f_M{}^3} \\ &= 3\left(\frac{\Delta\omega}{\omega_M}\right)^2\frac{A^2P_f}{N_0f_M} \qquad (3.37) \\ &= 3\beta^2\left(\frac{\text{S}}{N}\right)_{\text{SSB, DSB}}\end{aligned}$$

Recalling that, for large values of β, the bandwidth occupancy of the modulated signal is approximately $2\Delta f = 2\beta f_M$ and that the DSB bandwidth occupancy is $2f_M$, β may be interpreted as the ratio of the bandwidth needed with FM to that necessary for conventional AM or DSB transmission. We have shown that the signal-to-noise ratio improvement in FM is proportional to the square of this bandwidth multiplication factor β. Consequently, FM provides a means for increasing the bandwidth to obtain improved performance. Since increasing the FM bandwidth by β achieves the same results as increasing the signal power by β^2, FM may also be regarded as a method of exchanging power for bandwidth to keep the same performance.

The analysis of the signal-to-noise performances of PM follows along the same lines as that for FM, except that instead of the signal $y(t)$ of figure 3.10, we are now interested in its integral. That is, since $\theta_2(t) \approx \theta_1(t) + [n_1(t)/A]$ where, in this case, $\theta_1(t) = \Delta\theta f(t)$, it follows that $\theta_2(t)$ is the quantity of interest, not its derivative $y(t)$. Thus the input to the VCO must also be passed through an integrator in order to yield the desired output. The output signal power is clearly $(\Delta\theta)^2P_f$ while the noise power is $(1/A^2)N_0f_M$ resulting in a signal-to-noise ratio

$$\left(\frac{S}{N}\right)_{\text{PM}} = (\Delta\theta)^2\,\frac{A^2P_f}{N_0f_M} = (\Delta\theta)^2\left(\frac{S}{N}\right)_{\text{SSB, DSB}} \qquad (3.38)$$

In discussing phase-locked loop demodulation of FM (and PM), we have made some assumptions which should be emphasized. In particular, it was assumed that conditions were such that the VCO phase output was sufficiently close to the input phase that the approximation $\sin[\theta_1(t) - \theta_2(t)] \approx \theta_1(t) - \theta_2(t)$ was valid. However, the loop dynamics require that $\theta_2(t) \approx \theta_1(t) + [n_1(t)/A]$. Clearly, if the term $n_1(t)/A$ represents a phase angle of, say, more than 10°, then this approximation becomes unacceptable. But since

$$<\frac{n_1(t)}{A}> = 0$$

and

$$<\frac{n_1{}^2(t)}{A^2}> = \frac{N_0f_M}{A_2} \qquad (3.39)$$

(see footnote 4) it follows that $n_1(t)/A$ will be small compared to 10° only so long as the term N_0B/A^2 is sufficiently small. As N_0B/A^2 becomes large, it will not infrequently happen that this noise term causes $\theta_2(t)$ to be different enough from $\theta_1(t)$ so that the loop can no longer track the input. We have not specified the form of the filter $h(t)$. If possible, it is to be chosen so that $\theta_2(t) \approx \theta_1(t)$ regardless of the variation of $\theta_1(t)$, even in the presence of the noise $n_1(t)$. Techniques are available for mathematically specifying the optimum filter when the signal and noise spectra are known. Nevertheless, if the normalized noise power N_0B/A^2 is large, the difficulties mentioned above remain, regardless of the filter $h(t)$.

This *threshold* effect when the noise becomes sufficiently large is characteristic of any FM demodulating scheme. This may be seen intuitively by referring to figure 3.11. A frequency-modulated signal is shown in figure 3.11(a). The same signal is shown in figure 3.11(b) as perturbed by a small amount of additive noise, and in figure 3.11(c) as altered by noise with a considerably greater power. Since the information is conveyed in an FM signal by the instantaneous frequency, a measure of the effect of the noise exists in the comparison of the position of the zero crossings before and after the addition of the noise. It is seen that, as the noise increases, some zero crossings will be added by the noise while others will be eliminated entirely. When the noise reaches a level at which these phenomena become relatively common, the demodulated signal rapidly deteriorates.

SOME APPLICATIONS OF PHASE-LOCKED LOOPS

Before concluding this chapter, it is well to remark that phase-locked loops have many applications other than FM or PM demodulation. Some of these will be discussed subsequently; others have already been mentioned. In particular, in the case of DSB and SSB modulation it was suggested that the suppressed carrier be tracked by a phase-locked loop in order to acquire the reasonably accurate estimate of it which is necessary for product demodulation. The analysis of the phase-locked loop in this situation is identical to that presented in the previous section with two exceptions: First, the phase of the received signal $\theta_1(t)$ is constant except for a small variation caused by instabilities in the transmitter oscillator, by movement of the transmitter relative to the receiver, and perhaps by random fluctuations caused by the

[4] Since the effective loop bandwidth is f_M cps and it is only the noise within this bandwidth that can have any effect

$$<n_1^2(t)> = \lim \frac{1}{2T}\int_{-T}^{T} n_1^2(t)\ \mathrm{d}t = \phi_{n_1}(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_{n_1}(\omega)\ \mathrm{d}\omega$$
$$= \frac{N_0}{2}\int_{-f_M}^{f_M} \mathrm{d}f = N_0 f_M$$

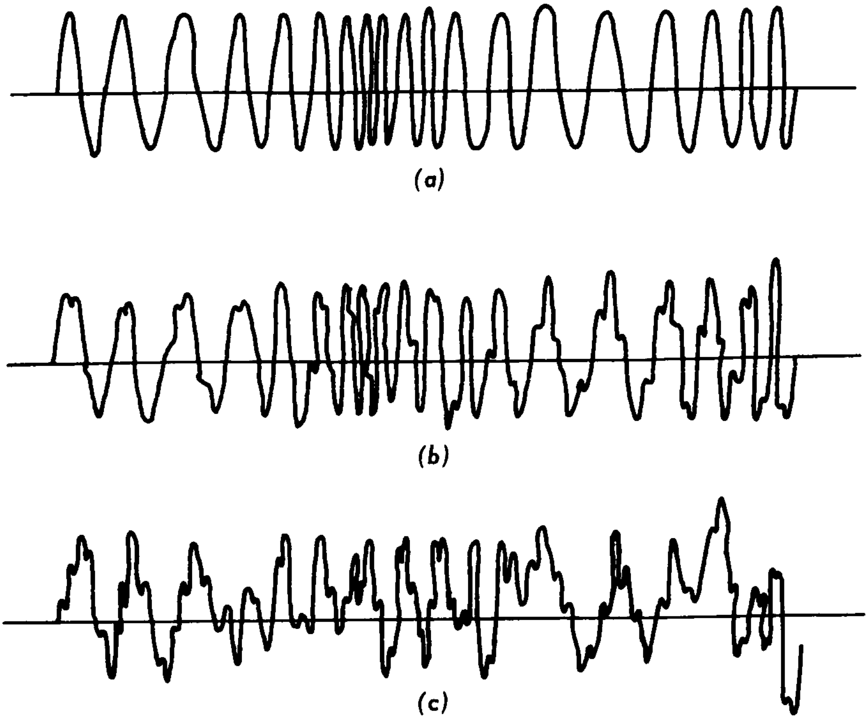


FIGURE 3.11—Effect of noise on a frequency-modulated wave. (a) A frequency-modulated wave (b) with small additive noise; (c) with large additive noise.

transmission medium. It is not caused to vary deliberately, and hence the bandwidth of $\theta_1(t)$ is very much less here than in the case of FM demodulation. Second, the desired signal output is not $y(t)$ but rather $\hat{x}(t)$, since it is the carrier itself, not just its phase which is to be estimated. The phase error of the estimate $\hat{x}(t)$, it was seen, is just $n_1(t)/a$ (where $\sqrt{2}a$ is the carrier amplitude) and represents an effective phase error power N_0B_L/a^2 (where B_L is the loop bandwidth). Thus, since the loop bandwidth B_L can be very narrow, the phase error can be reasonably small, even for quite small values of a. Since the phase error power N_0B_L/a^2 is the expected value of the square of the phase error, the square root of this quantity gives an estimate of the magnitude of the phase error which is encountered. By requiring $\sqrt{N_0B/a^2}<1/6$ radians for example, one can be reasonably sure that the phase error remains within tolerable limits. It will be recalled from "Noise Analysis of Amplitude Modulation Communication" in this chapter that

$$\left(\frac{S}{N}\right)_{AM}=\frac{P_T}{N_0f_M}$$

where P_T is the total power in the received signal; N_0, the noise spectral density; and f_M, the signal bandwidth. This was true under the assumption that the ratio of the power in the carrier to that in the modulation was negligibly small. Suppose, as an example, that it is required that the output signal-to-noise ratio $(S/N)_{\text{AM}}$ must be at least 1; that is, the signal power must be at least as great as the noise power, a generally quite marginal condition. Then

$$\frac{N_0B}{a^2}\left(\frac{S}{N}\right)_{\text{AM}}=\frac{N_0B}{a^2}\,\frac{P_T}{N_0f_M}=\frac{1}{36}$$

and

$$\frac{P_T}{a^2}=\frac{1}{36}\,\frac{f_M}{B}$$

Typically $f_M=6000$ cps, while the effective loop bandwidth of the carrier tracking loop can be made 1.0 cps or less. Thus

$$\frac{P_T}{a^2}=\frac{1}{6}\times 10^3$$

and, indeed, the required carrier power is negligible.

REFERENCES

1. Jahnke, Eugen; and Emde, Fritz: Tables of Functions With Formulae and Curves. Dover Publications, Inc., 1945.

2. Schwartz, Mischa: Information Transmission, Modulation and Noise. McGraw-Hill Book Co., Inc., 1959.

CHAPTER 4

Pulse Modulation

IN THIS CHAPTER we consider modulation techniques which differ fundamentally from those of the previous chapter. As a point of departure we begin with a discussion of the sampling theorem.

THE SAMPLING THEOREM

It is rather obvious that many of the measurements which are monitored on a spacecraft do not need to be observed continually. Temperatures, for example, generally vary quite slowly and readings need be taken only every minute or even every hour, certainly not continuously. It is perhaps surprising that not only temperature but any continuous time function can be observed only periodically *without any loss of information* concerning the original signal, whenever the signal power spectrum is identically zero for frequencies greater than some finite frequency W. The complete time function can be reconstructed from the periodic samples alone.

To verify this, consider the periodic function $S(t)$ illustrated in figure 4.1:

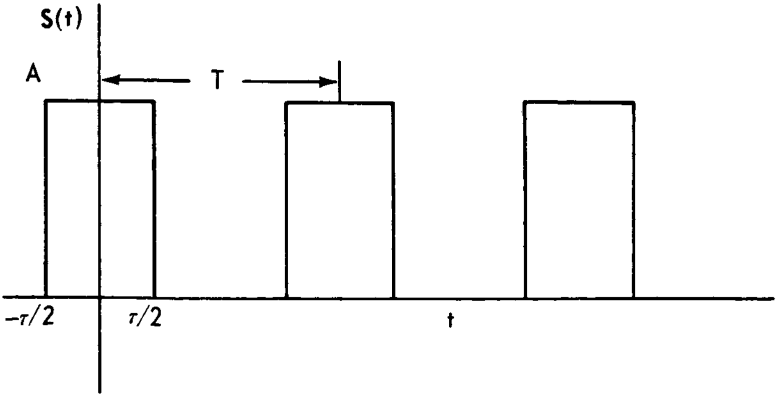


FIGURE 4.1—The periodic function $S(t)$.

Since $S(t)$ is periodic, it can be expanded in a Fourier series

$$S(t)=\frac{1}{T}\sum_{n=-\infty}^{\infty} c_n e^{j\omega_n t} \qquad \omega_n=\frac{2\pi n}{T} \tag{4.1}$$

where

$$c_n=A\tau\frac{\sin \omega_n\tau/2}{\omega_n\tau/2}$$

Now suppose we have some time function $x(t)$ which we wish to transmit. Consider the product $x(t)S(t)$. It will be recalled from "Expectation and Independence" in chapter 1 that

$$\begin{aligned}<x(t)x(t+\tau)S(t)S(t+\tau)> &= R_x(\tau)<S(t)S(t+\tau)>\\ &= R_x(\tau)\sum_{n=-\infty}^{\infty}\left|\frac{c_n}{T}\right|^2 e^{j\omega_n\tau}\end{aligned} \tag{4.2}$$

and, hence, upon taking the Fourier transform

$$\Phi(\omega)=\sum_{n=-\infty}^{\infty}\left|\frac{c_n}{T}\right|^2\Phi_x\left(\omega-\frac{2\pi n}{T}\right)$$

where $\Phi(\omega)$ is the power spectrum of the product and $\Phi_x(\omega)$ is the spectrum of the process $x(t)$.

Thus, if $x(t)$ has the spectrum shown in figure 4.2(a), the spectrum of the product $x(t)S(t)$ is that shown in figure 4.2(b). Note that so long as $W<1/2T$, there is no overlap of the components of the spectrum $\Phi(\omega)$.

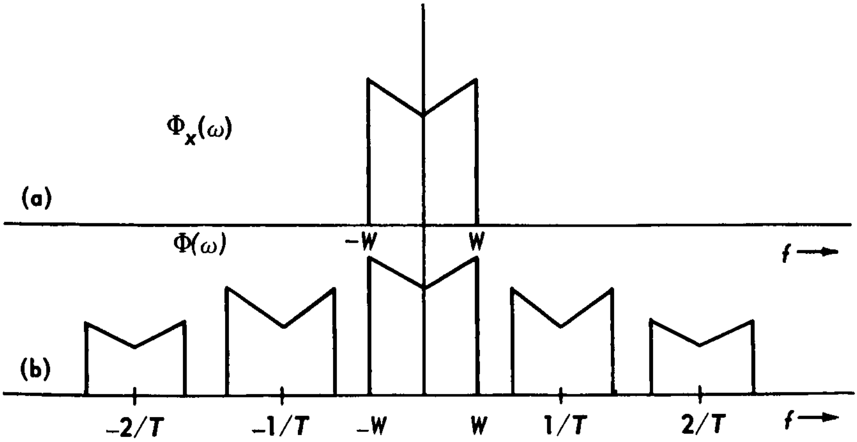


FIGURE 4.2—Power spectra. (a) $x(t)$; (b) $x(t)S(t)$.

Thus, the higher components can be filtered off, leaving just the original spectrum $\Phi_x(\omega)$ multiplied by the constant $|c_0/T|^2=(A\tau/T)^2$. Consequently, the time function $x(t)S(t)$ contains all of the information of the signal $x(t)$. Note that this is true regardless of the value of τ so long as $A\tau/T>0$. In particular, if we let $\tau \to 0$ and $A \to \infty$ such that $A\tau/T=1$, $S(t)$ becomes a periodic sequence of delta functions with amplitudes T, and $x(t)S(t)$ becomes a sequence of delta functions with amplitudes $Tx(nT)$. Thus all of the information needed to reconstruct the signal $x(t)$, when the spectrum of $x(t)$ is limited to W cps, is contained in the values of the amplitude of $x(t)$ at the instants of time $x(nT)$, subject only to the constraint that $T<1/2W$.

This result is known as the *sampling theorem.* It states that we need only concern ourselves with periodic samples of a bandwidth-limited time function. Only the sequence of numerical values $x(nT)$ need be transmitted. The complete function $x(t)$ may be reconstructed at the receiver by generating a series of delta functions of area $x(nT)$ and passing them through a lowpass filter.

TIME AND FREQUENCY MULTIPLEXING

There are a number of advantages associated with sampled data telemetry systems. First, some rather elegant techniques have been devised for transmitting sampled data. These methods at the same time are relatively easily implemented and achieve large signal-to-noise ratio increases at the expense of bandwidth.

In addition, it is generally considerably easier and more efficient to handle sampled data than continuous data. Typically, a spacecraft may contain 100 to 1000 data sources. Some method must be used to keep the information from each source separate. One method for doing this, called frequency multiplexing, consists of forming the products $x_i(t)$ $\sin \omega_i t$ for each data signal $x_i(t)$, $i=1, 2, \ldots$. The frequencies $\omega_i/2\pi$ must be such that the spectra of each of the signals do not overlap as shown in figure 4.3. The signal

$$x(t)=\sum_{i=1}^{N} x_i(t) \sin \omega_i(t) \tag{4.3}$$

then has the composite spectrum of figure 4.3, and $x(t)$ may be treated as a single source with a bandwidth

$$W=W_0+2\sum_{i=1}^{N} W_i \tag{4.4}$$

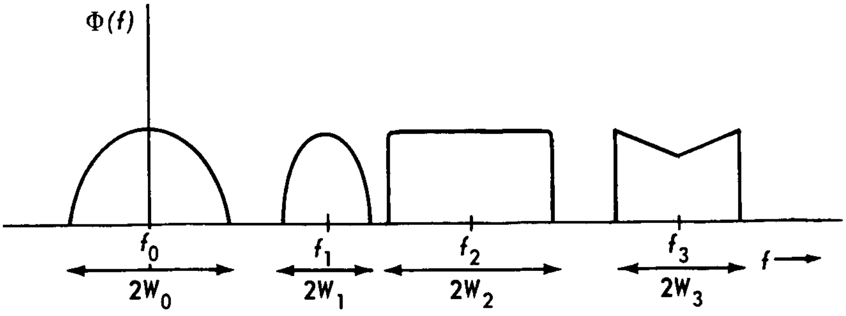


FIGURE 4.3—Frequency multiplexing.

Since the individual spectra do not overlap, the different signals $x_i(t)$ can be reconstructed at the receiver by proper filtering. Unfortunately, this method demands that each source be followed by a device for forming the product $x_i(t) \sin \omega_i t$, a procedure which is quite inefficient, particularly on board a spacecraft.

The alternative is to sample each of the signals $x_i(t)$, represent the samples as pulses of duration T/N where T is the sampling rate (here it is assumed that all signals have the same bandwidth so that $T = 1/2W$ is the same for all $x_i(t)$), and *time-multiplex* these samples as shown in figure 4.4.

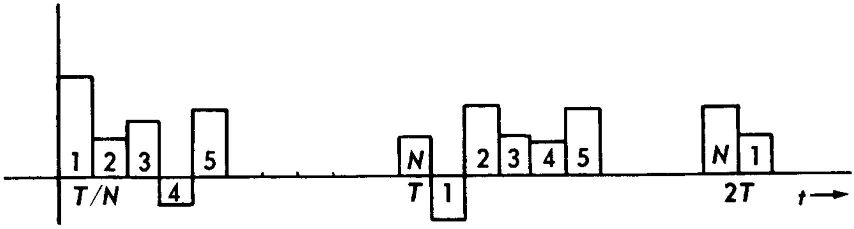


FIGURE 4.4—Time multiplexing.

The pulse labeled i corresponds to a sample of the process $x_i(t)$. If the bandwidths of the pulses are not the same, the sampling rates must be different, or all rates must be equal to that demanded by the signal with the largest bandwidth. Different sampling rates can readily be accommodated so long as they are integrally related. That is, suppose $x_1(t)$ has a bandwidth which is twice as great as $x_2(t)$ and the two are to be time multiplexed. Since $x_1(t)$ must be sampled twice as often as $x_2(t)$, they can easily be multiplexed as shown in figure 4.5. Thus, in time T two samples of $x_1(t)$ are transmitted while one sample of $x_2(t)$ is transmitted; both are sampled at the correct rate.

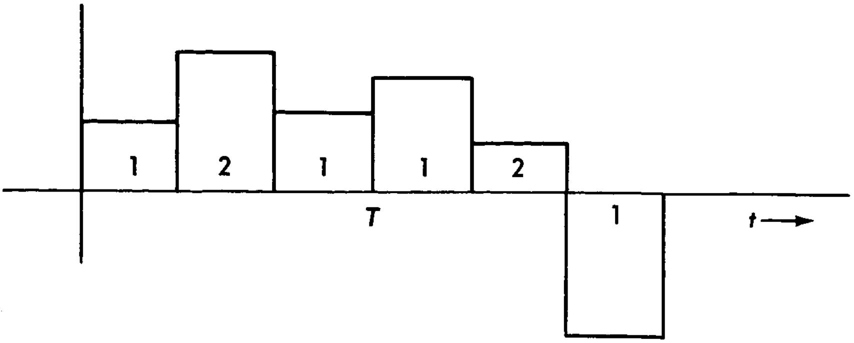


FIGURE 4.5—Time multiplexing of signals with unequal bandwidths.

A time-multiplexing system involves only the problem of *commutation* or the interspacing of samples from the various sources at the proper rate. No power-consuming auxiliary equipment, other than a minimum amount of switching devices, is required.

It might be supposed that, since each data signal is only being observed for an infinitesimal fraction of the time, the bandwidth requirements could be considerably reduced. Actually this is not the case as can be seen in the discussion of particular pulse-modulation systems. To see this heuristically, recall from "Bandwidth" in chapter 1 that the effective pulse width and effective bandwidth of a signal were related by $\tau = 1/2B$. The sampling theorem states that a signal of bandwidth W must be sampled at least every $T = 1/2W$ seconds. If this amplitude were transmitted as a pulse, the pulse could last only T seconds before the next sample must be sent. Thus the pulse width cannot be greater than $\tau = T = 1/2W$ and, hence, the effective bandwidth occupancy is $B = 1/2\tau = W$, the bandwidth of the signal. The same comment applies to multiplexed signals; frequency multiplexed signals require a bandwidth at least as great as the sum of the bandwidth of the individual signal

$$W = \sum_i W_i \tag{4.5}$$

(Actually, it is seen from equation (4.4) that the bandwidth there is about twice this value. However, it will be recalled from "Double- and Single-Sideband Modulation" in ch. 3 that half the bandwidth may be deleted without any loss of information. Thus, in the case of *single-sideband frequency multiplexing*, the above statement holds. This could be done, of course, only at the expense of additional equipment.) A time-multiplexed system involving N data sources has only T/N seconds per pulse as was seen in figure 4.4. Thus, the effective bandwidth is increased by a factor of N in the case of equal signal bandwidths $W_i = W$ ($\tau = 1/2NW$, $B = NW$). Similarly, the frequency-multiplexed signal bandwidth is increased by the same factor under the same condition as seen from equation (4.5). If the bandwidths are not equal, the frequency-multiplexed signals can be spaced more efficiently, in general, since there is no necessity for the sampling rates to be integrally related. However, this advantage is offset by the necessity of single-sideband multiplexing to avoid increasing the bandwidth by a factor of 2.

PULSE-MODULATION SYSTEMS AND MATCHED FILTERING

In the sections that follow, a number of pulse-modulation methods are discussed. To simplify the discussion, it will be assumed that the data input to the transmitter consists of a sequence of samples at some average rate, say R samples per second. Thus each sample has $T=1/R$ seconds in which to be transmitted. It is unimportant whether this sequence comes from one source or is the time-multiplexed output from a number of sources. A pulse-modulation system involves the transmission of a particular waveform $f(t)$ representing the sample in question for a period of time T seconds. The transmitted signal, therefore, is allowed to change form only every T seconds.

Before proceeding to discuss various pulse-modulating schemes in more detail, it is of interest to consider the generic form of the demodulators for pulse modulation. The pulse-modulated signal, as observed, is characterized by a waveform $f(t)$ which is transmitted without change for $\nu T<t<(\nu+1)T$. To prescribe the desired demodulator, it is necessary to specify its exact function. Let us assume first that the demodulator network is to be linear. (Surprisingly, for the situations of interest here, the linearity constraint can be omitted without altering the conclusions.) We would like the output of the network to be large when $f(t)$ has been transmitted and small when it has not. We are, in fact, interested only in whether $f(t)$ was actually transmitted, not in reproducing it at the receiver, since presumably we know the functional form of $f(t)$. Thus, since we want to decide whether or not $Af(t)$ (A is a function of the transmitted power and of the channel attenuation factor k) has been received over the time interval $\nu T<t<(\nu+1)T$, the logical time to observe the output of the network is at the instant of time $(\nu+1)T$ after all the pertinent information has been received. At this time we would like the output to be as large as possible if $Af(t)$ were present and as small as possible otherwise. Since noise $n(t)$ is always present, the output due to the noise, in particular, should be kept small. Thus, if $Ag[(\nu+1)T]$ is the output due to the signal $Af(t)$ at the time $(\nu+1)T$ and if the average noise power output is N_H, we would like to find the linear system which maximizes the ratio $A|g[(\nu+1)T]|/N_H$. Since $A^2g^2(t)$ is a monotonic function of $A|g(t)|$, it is equivalent to maximizing the ratio $A^2g^2(\nu+1)T/N_H$. Finally, designating by N_0 the single-sided input noise spectral density and letting

$$A^2E_f = A^2\int_{-\infty}^{\infty} f^2(t)\,\mathrm{d}t = \frac{A^2}{2\pi}\int_{-\infty}^{\infty} |F(j\omega)|^2\,\mathrm{d}\omega \tag{4.6}$$

be the received signal energy, the expression $A^2g^2[(\nu+1)T]/N_H$ is maximized if and only if the ratio $(N_0/2)g^2[(\nu+1)T]/N_HE_f$ is maximized, since E_f and $N_0/2$ are constant, independent of the network at the receiver. Let us denote the impulse response of the linear system by $h(t)$ and its transfer function by $H(j\omega)$. Then, from "Linear Systems" in chapter 1

$$g(t) = \int_{-\infty}^{\infty} H(j\omega)F(j\omega)\mathrm{e}^{j\omega t}\,\frac{\mathrm{d}\omega}{2\pi} \tag{4.7}$$

where $F(j\omega)$ is the Fourier transform of $f(t)$, and

$$g[(\nu+1)T] = \int_{-\infty}^{\infty} H(j\omega)F(j\omega)\mathrm{e}^{j\omega(\nu+1)T}\,\frac{\mathrm{d}\omega}{2\pi} \tag{4.8}$$

Assuming the noise spectral density is flat over the region of interest with the two-sided spectral density $N_0/2$

$$N_H = \frac{N_0}{2}\int_{-\infty}^{\infty} |H(j\omega)|^2\,\frac{\mathrm{d}\omega}{2\pi} \tag{4.9}$$

(see footnote 5). It thus becomes our goal to maximize the ratio

$$\frac{\frac{N_0}{2}\,g^2[(\nu+1)T]}{N_HE_f} = \frac{\left|\int_{-\infty}^{\infty} H(j\omega)F(j\omega)\mathrm{e}^{j\omega(\nu+1)T}\,\frac{\mathrm{d}\omega}{2\pi}\right|^2}{\int_{-\infty}^{\infty} |H(j\omega)|^2\,\frac{\mathrm{d}\omega}{2\pi}\int_{-\infty}^{\infty} |F(j\omega)|^2\,\frac{\mathrm{d}\omega}{2\pi}} \tag{4.10}$$

[5] See "Linear Systems" and "Spectra and Autocorrelation" in ch. 1. Since

$$N_H(t) = \int_{-\infty}^{\infty} n(t-\tau)h(\tau)\,\mathrm{d}\tau$$

the output noise power is

$$<N_H^2(t)> = <\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} n(t-\tau)n(t-\eta)h(\tau)h(\eta)\,\mathrm{d}\tau\,\mathrm{d}\eta>$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} <n(t-\tau)n(t-\eta)>h(\tau)h(\eta)\,\mathrm{d}\tau\,\mathrm{d}\eta$$

This last step follows from the fact that we are taking the expectation with respect to the time t, and $h(\tau)$ and $h(\eta)$ are independent of t. But $<n(t-\tau)n(t-\eta)> = (N_0/2)\delta(\tau-\eta)$ and so

$$N_H = <n_H^2(t)> = \frac{N_0}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau)h(\eta)\delta(\tau-\eta)\,\mathrm{d}\tau\,\mathrm{d}\eta$$

$$= \frac{N_0}{2}\int_{-\infty}^{\infty} h^2(\tau)\,\mathrm{d}\tau = \frac{N_0}{2}\int_{-\infty}^{\infty} |H(j\omega)|^2\,\frac{\mathrm{d}\omega}{2\pi}$$

We can easily obtain a bound on this ratio by using the well-known Schwarz's inequality which states that

$$\left|\int_{-\infty}^{\infty} X(s)Y(s)\,\mathrm{d}s\right|^2 \leqq \int_{-\infty}^{\infty} |X(s)|^2\,\mathrm{d}s \int_{-\infty}^{\infty} |Y(s)|^2\,\mathrm{d}s \tag{4.11}$$

Thus, since

$$\left|F(j\omega)\mathrm{e}^{j\omega(\nu+1)T}\right| = \left|F(j\omega)\right|$$

it follows from equation (4.10) that

$$\frac{\frac{N_0}{2}[g^2(\nu+1)T]}{N_H E_f} \leq 1 \tag{4.12}$$

Suppose, now, we let

$$H(j\omega)\mathrm{e}^{j\omega(\nu+1)T} = F(-j\omega) \tag{4.13}$$

Then

$$\frac{\frac{N_0}{2}g^2[(\nu+1)T]}{N_H E_f} = \frac{\left[\int_{-\infty}^{\infty} |F(j\omega)|^2 \frac{\mathrm{d}\omega}{2\pi}\right]^2}{\int_{-\infty}^{\infty} |F(j\omega)|^2 \frac{\mathrm{d}\omega}{2\pi} \int_{-\infty}^{\infty} |F(j\omega)|^2 \frac{\mathrm{d}\omega}{2\pi}} = 1$$

Thus, the upper bound on the ratio (4.10) is achieved if $H(j\omega) = F(-j\omega)\mathrm{e}^{-j\omega(\nu+1)T}$ or if

$$\begin{aligned} h(t) &= \int_{-\infty}^{\infty} H(j\omega)\mathrm{e}^{j\omega t}\frac{\mathrm{d}\omega}{2\pi} \\ &= \int_{-\infty}^{\infty} F(-j\omega)\mathrm{e}^{-j\omega[(\nu+1)T-t]}\,\frac{\mathrm{d}\omega}{2\pi} \\ &= \int_{-\infty}^{\infty} F(j\omega)\mathrm{e}^{j\omega[(\nu+1)T-t]}\,\frac{\mathrm{d}\omega}{2\pi} \\ &= f[(\nu+1)T-t] \end{aligned} \tag{4.14}$$

and the network with the impulse response $h(t) = f[(\nu+1)T-t]$ is optimum in the sense described.

Let $y(t)$ be the actual received signal. Then the optimum test to determine whether $y(t) = Af(t)+n(t)$, where $n(t)$ is white additive noise and A is some arbitrary gain constant, is to pass it through the network with the impulse response (4.14). Since, from "Linear Systems" in chapter 1, the output from such a network due to the input $y(t)$ is

$$
\begin{aligned}
g[(\nu+1)T] &= \int_{-\infty}^{\infty} h[(\nu+1)T-\tau]y(\tau)\,d\tau \\
&= \int_{-\infty}^{\infty} f(\tau)y(\tau)\,d\tau \\
&= \int_{\nu T}^{(\nu+1)T} f(\tau)y(\tau)\,d\tau
\end{aligned}
\tag{4.15}
$$

where the last step follows from the fact that $f(t)$ lasts only T seconds, $\nu T<t<(\nu+1)T$. The optimum pulse-modulation receiver, it would seem, forms the integrals

$$g_i[(\nu+1)T]=\int_{\nu T}^{(\nu+1)T} f_i(\tau)y(\tau)\,d\tau \tag{4.16}$$

for all possible signals $f_i(t)$ and selects the largest as that corresponding to the signal that was actually transmitted. This conclusion follows from the fact that, if the signal $f_m(t)$ is transmitted, the output $g_m[(\nu+1)T]$ should be large, whereas all other outputs g_i $[(\nu+1)T]$, where $i\neq m$, should be small. That such a receiver, consisting of a bank of detectors forming the quantities $g_i[(\nu+1)T]$ (such detectors are called *correlation detectors* or *matched filters* for rather obvious reasons), is in fact optimum can be proved rigorously when the additive noise is white and Gaussian. The optimum decision, when each signal $f_i(t)$ has the same energy, is, indeed, to select the maximum over i of the outputs $g_i[(\nu+1)T]$. When the signal energies are not independent of i, the optimum decision is to select the largest of the terms

$$g_i[(\nu+1)T]-\frac{A}{2}E_{f_i} \tag{4.17}$$

where

$$E_{f_i}=\int_{\nu T}^{(\nu+1)T} f_i^2(t)\,dt$$

is the energy in the ith signal. The same procedure is, of course, repeated for every time interval $\nu T<t<(\nu+1)T$ for all integers ν.

PULSE AMPLITUDE MODULATION (PAM)

Perhaps the most obvious method for transmitting sampled data is pulse amplitude modulation. If the data sample is x_ν the signal $\sqrt{2}x_\nu B$ sin $\omega_c t$ is transmitted, $\nu T<t<(\nu+)T$, the received signal then becoming $y(t)=\sqrt{2}Ax_\nu \sin \omega_c t+n(t)$ (see footnote 6). As discussed in the previous section, the optimum detector forms the quantity

[6] There will be, in general, a phase and probably even a frequency shift between the transmitter and the receiver. However, this will cause no difficulty so long as the received phase and frequency are determined at the receiver and used to generate the local signals $f_i(t)$. This knowledge will be assumed here so that the phase and frequency shift can safely be ignored.

$$g_i[(\nu+1)T]-\frac{A}{2}E_{f_i}=\int_{\nu T}^{(\nu+1)T} y(t)\sqrt{2}x_i \sin\omega_c t\, dt$$
$$-\frac{A}{2}\int_{\nu T}^{(\nu+1)T} 2x_i^2 \sin^2\omega_c t\, dt \quad (4.18)$$

for every possible amplitude x_i of the received signal, and selects the largest of these as the best estimate of the received signal. But a condition that the quantity $g_i[(\nu+1)T]$ be a maximum is that

$$\frac{d}{dx_i} g_i[(\nu+1)T]=0$$

or that

$$\int_{\nu T}^{(\nu+1)T} y(t)\sqrt{2}\sin\omega_c t\, dt = x_i A\int_{\nu T}^{(\nu+1)T} 2\sin^2\omega_c t\, dt$$
$$=x_i AT \quad (4.19)$$

where the carrier frequency f_c has been chosen to be some multiple of half the reciprocal of the pulse period $T, f_c=\omega_c/2\pi=k/2T$ for some integer k. The optimum estimate $\hat{x}_\nu$ of the amplitude of the received signal, then is

$$\hat{x}_\nu=\frac{1}{AT}\int_{\nu T}^{(\nu+1)T} y(t)\sqrt{2}\sin\omega_c t\, dt \quad (4.20)$$

and the receiver is simply that illustrated schematically in figure 4.6

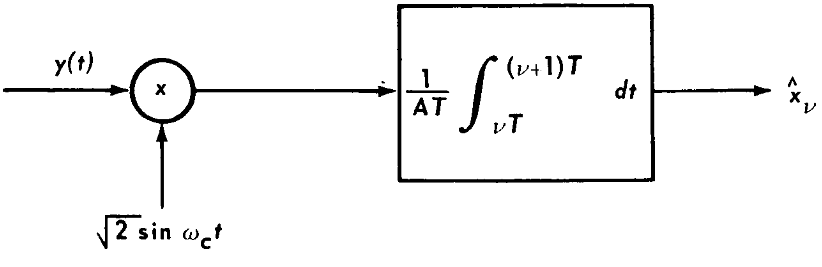


FIGURE 4.6—A PAM detector.

To determine the signal-to-noise ratio at the output of a PAM detector, we observe that the output noise power [5] is given by

$$N_H=\frac{N_0}{2}\int_{-\infty}^{\infty} h^2(\tau)\, d\tau$$

In this case, we have, from equation (4.20),

$$\hat{x}_\nu = \frac{1}{AT}\int_{\nu T}^{(\nu+1)T} y(t)\sqrt{2}\,\sin\,\omega_c t\;\mathrm{d}t$$

$$= \int_{\nu T}^{(\nu+1)T} y(t)h[(\nu+1)T-t]\;\mathrm{d}t \tag{4.21}$$

Hence

$$h[(\nu+1)T-t] = \frac{\sqrt{2}}{AT}\sin\,\omega_c t \qquad \nu T<t<(\nu+1)T$$

or

$$h(\tau) = \frac{\sqrt{2}}{AT}\sin\,\omega_c[(\nu+1)T-\tau] \qquad 0<\tau<T$$

and consequently

$$N_H = \frac{N_0}{2}\int_0^T \frac{2\,\sin^2\,\omega_c[(\nu+1)T-\tau]}{A^2T^2}\,\mathrm{d}\tau = \frac{N_0}{2A^2T} \tag{4.22}$$

The output *signal* is that part of the output produced by the input signal only, so that, replacing $y(t)$ with $\sqrt{2}Ax_\nu\,\sin\,\omega_c t$, in equation (4.20),

$$S_0[(\nu+1)T] \equiv \frac{1}{AT}\int_{\nu T}^{(\nu+1)T} Ax_\nu\sqrt{2}\,\sin^2\,\omega_c t\;\mathrm{d}t = x_\nu \tag{4.23}$$

The output signal power is defined as

$$P_{\text{ave}} = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} S_0^2(t)\;\mathrm{d}t = E[S_0^2] = E\{S_0^2[(\nu+1)T]\} \tag{4.24}$$

The last step follows from the fact that $S_0(t)$ is assumed to be stationary. Hence, $E[S_0^2(t)]$ is independent of time and may be evaluated at any particular instant of time. From equation (4.23) we have

$$E\{S_0^2[(\nu+1)T]\} = E(x_\nu^2) \tag{4.25}$$

and, recalling from "Expectation and Independence" in chapter 1 that

$$E(x_\nu^2) = \int_{-\infty}^{\infty} x_\nu^2 p(x_\nu)\;\mathrm{d}x$$

it is only necessary to know the probability density function of the signal x_ν in order to determine the signal power. For purpose of illustration it will be assumed that

$$p(x_\nu)=\begin{cases}\dfrac{1}{a} & -\dfrac{a}{2}<x_\nu<\dfrac{a}{2}\\[2ex] 0 & \text{Otherwise}\end{cases}\tag{4.26}$$

so that x_ν is equally likely to assume any value between $-a/2$ and $a/2$ (see footnote 7). In this case

$$E(x_\nu^2)=\frac{1}{a}\int_{-a/2}^{a/2}x_\nu^2\,dx_\nu=\frac{a^2}{12}\tag{4.27}$$

Thus, the output signal-to-noise ratio is

$$\left(\frac{S}{N}\right)_{\text{PAM}}=\frac{P_{\text{ave}}}{N_H}=\frac{\dfrac{a^2}{12}A^2T}{\dfrac{N_0}{2}}\tag{4.28}$$

It may be shown by an analysis similar to that in "Bandwidth" in chapter 1 that the bandwidth of the PAM signal is proportional to $1/T$. An interesting measure of the effective bandwidth of a signal is afforded by asking the question: How far in frequency must two channels be separated if the *cross-modulation* between any two of them is to be kept to an insignificant level? Suppose, in fact, that a number of PAM channels were to be operated simultaneously at the carrier frequencies ω_i where $i=1, 2, \ldots.$ Then the effect of the *signal* in the ith channel on the output of the jth channel demodulator is simply

$$\begin{aligned}\int_{\nu T}^{(\nu+1)T} y_i(t)\sqrt{2}\sin\omega_j t\,dt&=\int_{\nu T}^{(\nu+1)T} x_\nu(i)2\sin\omega_i t\sin\omega_j t\,dt\\ &=x_\nu(i)\left[\int_{\nu T}^{(\nu+1)T}\cos(\omega_i-\omega_j)t\,dt\right.\\ &\qquad\left.+\int_{\nu T}^{(\nu+1)T}\cos(\omega_i+\omega_j)t\,dt\right]\end{aligned}\tag{4.29}$$

which is identically zero if ω_i, ω_j, and $\omega_i-\omega_j$ are all nonzero multiples of the term π/T. Thus, in order to keep the cross-modulation zero, it is necessary to separate the channels in frequency by an amount $f_i-f_j=k/2T$, for any value of $k=1, 2, \ldots.$ Thus the *effective* bandwidth of any channel is just $B=1/2T$ cps.

[7] If this is not the situation, it is often desirable to precondition the signal to render its distribution in the form of equation (4.26) (cf. ch. 5).

The noise spectral density is $N_0/2$ watts/cps and the average signal power at the *input* to the receiver is

$$P_s = A^2 \int_{-\infty}^{\infty} x_\nu^2 p(x_\nu)\, dx_\nu = A^2 \frac{a^2}{12} \tag{4.30}$$

Thus the output signal-to-noise ratio (equation (4.28)) can be expressed in terms of the input signal-to-noise ratio $(S/N)_i$ as follows:

$$\left(\frac{S}{N}\right)_0 = \frac{A^2 \frac{a^2}{12} T}{\frac{N_0}{2}} = \left(\frac{P_s}{N_0 B}\right) = \left(\frac{S}{N}\right)_i \tag{4.31}$$

Note that this is exactly the same relationship that was obtained for DSB and SSB modulation.

PHASE-SHIFT KEYED MODULATION

Another rather common pulse-modulation technique, called phase-shift keying, is to transmit the signal

$$\sqrt{2} \sin\, (\omega_c t + x_\nu) \qquad \nu T < t < (\nu+1)T \tag{4.32}$$

to convey the data x_ν. Here, of course, $0 < x_\nu < 2\pi$ so that there is no ambiguity at the receiver. Thus the phase, rather than the amplitude, conveys the information in a phase-shift keyed modulation (PSK) system. The advantage of this method over PAM rests in the fact that the amplitude of the signal remains constant. This is not an insignificant advantage in space telemetry, since transmitters which work at a constant amplitude are considerably more efficient than those which must produce variable amplitudes.

From "Pulse Modulation Systems and Matched Filtering" in this chapter, the optimum PSK receiver, since the received signal energy is now independent of x_ν, must form the integrals

$$\int_{\nu T}^{(\nu+1)T} y(t) \sqrt{2} \sin\, (\omega_c t + x_i)\, dt \tag{4.33}$$

for all values of $0 < x_i < 2\pi$ and select the largest. But this expression can be written

$$\cos x_i \int_{\nu T}^{(\nu+1)T} y(t)\sqrt{2}\,\sin\,\omega_c t\;\mathrm{d}t + \sin x_i \int_{\nu T}^{(\nu+1)T} y(t)\sqrt{2}\,\cos\,\omega_c t\;\mathrm{d}t$$

$$= X\cos x_i + Y\sin x_i \qquad (4.34)$$

where

$$X = \int_{\nu T}^{(\nu+1)T} \sqrt{2}y(t)\,\sin\,\omega_c t\;\mathrm{d}t$$

and

$$Y = \int_{\nu T}^{(\nu+1)T} \sqrt{2}y(t)\,\cos\,\omega_c t\;\mathrm{d}t$$

The maximum of these with respect to x_i must satisfy the condition that

$$\frac{\mathrm{d}}{\mathrm{d}x_i}\,(X\cos x_i + Y\sin x_i) = 0$$

or that

$$x_i = \tan^{-1}\frac{Y}{X} \qquad (4.35)$$

Thus the optimum estimate of x_ν is $\hat{x}_\nu = \tan^{-1}(Y/X)$ and the optimum receiver is that depicted in figure 4.7.

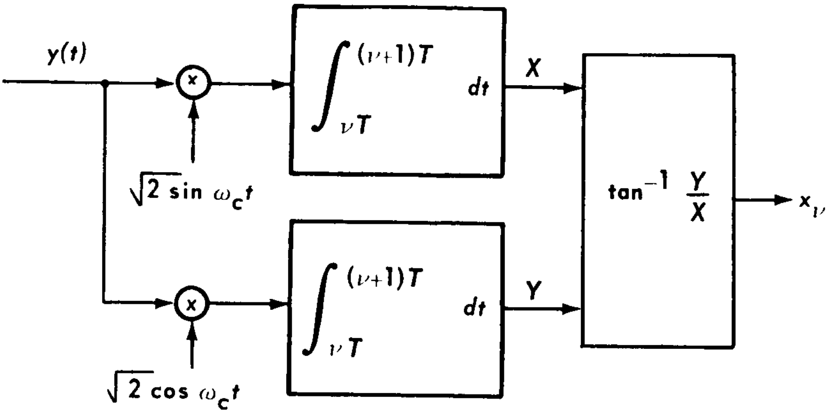


FIGURE 4.7—A PSK detector.

It is observed that in the absence of noise

$$y(t) = \sqrt{2}A\,\sin\,(\omega_c t + x_\nu)$$

and

$$\begin{aligned} X &= \int_{\nu T}^{(\nu+1)T} 2A \sin \omega_c t \sin (\omega_c t + x_\nu)\, dt \\ &= A \int_{\nu T}^{(\nu+1)T} [\cos x_\nu - \cos (2\omega_c t + x_\nu)]\, dt \\ &= AT \cos x_\nu \end{aligned}$$

while

$$\begin{aligned} Y &= \int_{\nu T}^{(\nu+1)T} 2A \cos \omega_c t \sin (\omega_c t + x_\nu)\, dt \\ &= A \int_{\nu T}^{(\nu+1)T} [\sin x_\nu + \sin (2\omega_c t + x_\nu)]\, dt \\ &= AT \sin x_\nu \end{aligned}$$

where it is again assumed that $\omega_c = \pi k/T$ for some integer k. Hence

$$\hat{x}_\nu = \tan^{-1} \frac{AT \sin x_\nu}{AT \cos x_\nu} = x_\nu$$

and the estimate of the signal is exact in the absence of noise. The analysis of the output signal-to-noise ratio for PSK is somewhat more involved than that for PAM and will not be carried out here. The results of such an analysis, however, would indicate performance approximately equal to that of PAM. In addition, it is easily verified that the same comments concerning the spectrum of a PAM signal as well as its effective bandwidth occupancy apply equally to a PSK-modulated signal.

PULSE DURATION MODULATION

Another rather interesting pulse-modulation scheme is pulse-duration modulation (PDM). Although there are a number of variations on this technique, only one PDM system will be discussed here. The data x_ν are normalized so that $0 < x_\nu < 1$ for all values of ν. The modulated signal then takes the form

$$\left.\begin{array}{ll} \sqrt{2} \sin \omega_c t & \nu T < t < (\nu + x_\nu) T = t_\nu \\ -\sqrt{2} \sin \omega_c t & t = (\nu + x_\nu) T < t < (\nu + 1) T \end{array}\right\} \qquad (4.36)$$

Since, again, the signal energy is independent of x_ν, the optimum detector, from "Pulse Modulation Systems and Matched Filtering" in this chapter, forms the integrals

$$\int_T^{t_i} y(t)\ 2\sin\omega_c t\ dt - \int_{t_i}^{(\nu+1)T} y(t)\ 2\sin\omega_c t\ dt \tag{4.37}$$

for all values of t_i where $\nu T < t_i < (\nu+1)T$.

Differentiating equation (4.37) establishes only the rather obvious condition that the maximum value of t_i satisfies the equation

$$y(t_i)\sin\omega_c t_i = 0$$

Unfortunately, there will, in general, be many values of t_i satisfying this equation, and a separate integration must be performed for each of these values. Clearly, such a receiver would generally not be practical.

A practical PDM receiver, however, does result if the value of x_ν is *quantized*. That is, if $i/N < x_\nu < (i+1)/N$, where $i=0, 1, 2, \ldots, N-1$, the value $x_\nu = [(i+1)/2]/N$ is transmitted. Then only the discrete values of $t_i = \nu T + [(i+1)/2](T/N)$ need be investigated. Further, since

$$\int_{\nu T}^{t_j} \sqrt{2}y(t)\sin\omega_c t\ dt$$
$$= \int_{\nu T}^{t_0} \sqrt{2}y(t)\sin\omega_c t\ dt + \sum_{i=1}^{j} \int_{t_{i-1}}^{t_i} \sqrt{2}y(t)\sin\omega_c t\ dt \tag{4.38}$$

the $N+1$ integrals

$$I_0 = \int_{\nu T}^{t_0} \sqrt{2}y(t)\sin\omega_c t\ dt$$
$$I_i = \int_{t_{i-1}}^{t_i} \sqrt{2}y(t)\sin\omega_c t\ dt \qquad i=1, \ldots, N-1 \tag{4.39}$$
$$I_N = \int_{t_N}^{(\nu+1)T} \sqrt{2}y(t)\sin\omega_c t\ dt$$

can be formed and the maximum over j, where $j=0, 1, \ldots, N-1$, of the summations

$$\sum_{i=0}^{j} I_i - \sum_{i=j+1}^{N} I_i \tag{4.40}$$

used to determine the estimate of the data x_ν. This system is illustrated in figure 4.8.

Note that only one integrator is necessary; the various quantities I_i can be determined, for example, by sampling the output of the integrator at each instant of time t_i and forming the difference

$$I_i = \int_{\nu T}^{t_i} \sqrt{2}y(t)\sin\omega_c t\ dt - \int_{\nu T}^{t_{i-1}} \sqrt{2}y(t)\sin\omega_c t\ dt \tag{4.41}$$

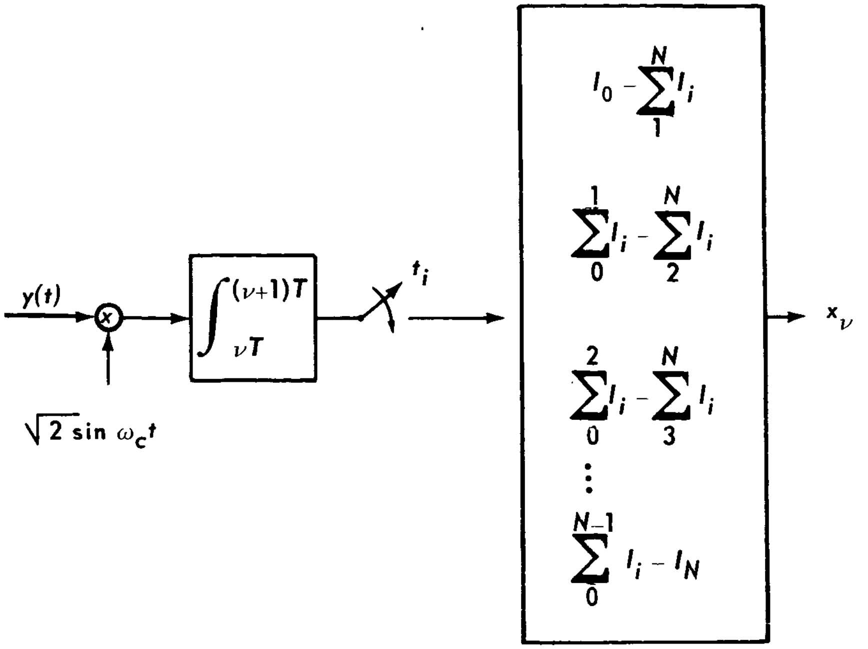


FIGURE 4.8—A discrete PDM detector.

The advantage of considering a discrete set of data values rather than a continuous set is that a simple receiver can be built. If t_i were continuous, a receiver analogous to that of figure 4.8 could be used by determining the value of t_i, not in advance but by the zero crossings of the signal $y(t)$. However, there could be quite a large number of zero crossings, and since this number would vary from pulse to pulse, the receiver would necessarily be considerably more complex. In addition, it is dubious that it would offer much advantage over the one described here.

The output signal-to-noise ratio is again somewhat difficult to derive. However, when the input signal-to-noise ratio is high, a rather interesting effect can be observed. The received signal has the form

$$y(t) = \pm\sqrt{2}A \sin \omega_c t + n(t) \qquad t_i < t < t_{i+1}$$

$$i = 0, 1, \ldots, N-1 \qquad (4.42)$$

where the sign remains the same over the interval in question. The integral I_i is then

$$I_i = \pm \int_{t_{i-1}}^{t_i} 2A \sin^2 \omega_c t \, dt + \int_{t_{i-1}}^{t_i} \sqrt{2} n(t) \sin \omega_c t \, dt \qquad (4.43)$$

The noise $n(t)$ is white and Gaussian with $E[n(t)]=0$ and $E[n(t)n(u)] = (N_0/2)\delta(t-u)$. If ω_c is chosen to be some multiple of $\pi/\Delta t$ where $\Delta t = t_i - t_{i-1}$, then, referring to "Gaussian Statistics" in chapter 1, equations (1.70) and (1.71), we see that

$$\mu_i = E(I_i) = \pm A\Delta t + \int_{t_{i-1}}^{t_i} E[n(t)]\, 2 \sin \omega_c t \, \mathrm{d}t = \pm A\Delta t \qquad (4.44)$$

and

$$\begin{aligned}
\sigma_i^2 &= E(I_i^2) - \mu_i^2 \\
&= \int_{t_{i-1}}^{t_i} 2\, E[n(t)n(\omega)] \sin \omega_c t \sin \omega_c u \, \mathrm{d}t \, \mathrm{d}u \\
&= \frac{N_0}{2} \int_{t_{i-1}}^{t_i} 2\delta(t-u) \sin \omega_c t \sin \omega_c u \, \mathrm{d}t \, \mathrm{d}u \\
&= \frac{N_0}{2} \int_{t_{i-1}}^{t_i} 2 \sin^2 \omega_c t \, \mathrm{d}t = \frac{N_0}{2} \Delta t \qquad \text{(see footnote 8)} \qquad (4.45)
\end{aligned}$$

Since $n(t)$ is a Gaussian process, then I_i is a Gaussian random variable (see "Gaussian Statistics" in ch. 1):

$$p(I_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(I_i - \mu_i)^2}{2\sigma_i^2} \right] \qquad (4.46)$$

and the probability that $I_i < 0$ given that $\mu_i = +A\Delta t$ is

$$\begin{aligned}
Pr(I_i < 0 | \mu_i = A\Delta t) &= \frac{1}{\sqrt{2\pi}\sigma_i} \int_{-\infty}^{0} \exp\left[-\frac{(I_i - \mu_i)^2}{2\sigma_i^2} \right] \mathrm{d}I_i \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mu_i/\sigma_i}^{\infty} \mathrm{e}^{-x^2/2} \mathrm{d}x \\
&= \frac{1}{2} \left\{ 1 - \operatorname{erf}\left[\left(\frac{A^2 \Delta t}{N_0/2} \right)^{1/2} \right] \right\} \qquad (4.47)
\end{aligned}$$

Similarly, the probability that $I_i > 0$ given that $\mu_i = -A\Delta t$ is

$$Pr[I_i > 0 | \mu_i = -A\Delta t] = \frac{1}{2} \left\{ 1 - \operatorname{erf}\left[\left(\frac{A^2 \Delta t}{N_0/2} \right)^{1/2} \right] \right\} \qquad (4.48)$$

[8] Actually, the integrals I_0 and I_N involve half intervals and both μ_i and σ_i^2 should be divided by one-half in these cases. However these "end effects" are insignificant for large values of N and will be neglected here. They could have been eliminated entirely by a slightly different quantization procedure.

When $(A^2\Delta t/N_0/2)$ is large, it is seen, again referring to "Gaussian Statistics" in chapter 1, that

$$\operatorname{erf}\left[\left(\frac{A^2\Delta t}{N_0/2}\right)^{1/2}\right]\approx 1$$

and hence that

$$Pr[I_i<0|\mu_i=A\Delta t]=Pr[I_i>0|\mu_i=-A\Delta t]\approx 0$$

Thus, the probability that a positive pulse of length Δt contributes a negative quantity at the output of the integrator or, conversely, that a negative pulse contributes a positive quantity, becomes negligibly small as $(A^2\Delta t)/(N_0/2)$ becomes large. (For practical purposes, this usually means that $(A^2\Delta t)/(N_0/2)>5$.) When this happens, an error in reception almost never occurs. Thus the difference between the received signal and the data is due solely to the inevitable error in quantizing the continuous signal x_ν into one of a discrete set of values. This quantization error is easily determined. The sign of the transmitted pulse changes at time t_i for any value of x_ν satisfying the inequalities $i/N<x_\nu<(i+1)/N$. If the distribution of x_ν is flat, $p(x_\nu)=1$, where $0<x_\nu<1$, then x_ν is equally likely to have any value in this range. Hence, the expected value of x_ν given that the pulse changed sign at time t_i is

$$E(x_\nu|t_i)=\int_{i/N}^{(i+1)/N} x_\nu p(x_\nu|t_i)\ \mathrm{d}x_\nu=N\int_{i/N}^{(i+1)/N} x_\nu\ \mathrm{d}x_\nu$$

$$=\frac{(i+\frac{1}{2})}{N} \qquad (4.49)$$

since

$$p(x_\nu|t_i)=\begin{cases} N \text{ (see footnote 9)} & i/N<x_\nu<\dfrac{i+1}{N} \\ 0 & \text{Otherwise}\end{cases}$$

In addition

$$\sigma_\nu^2=E(x_\nu^2|t_i)-E^2(x_\nu|t_i)$$

$$=N\int_{i/N}^{(i+1)/N} x_\nu^2\ \mathrm{d}x_\nu-\frac{(i+\frac{1}{2})^2}{N^2}=\frac{1}{12N^2}=\frac{(\Delta t)^2}{12T^2} \qquad (4.50)$$

The signal variance is

$$\sigma_s^2=E(x_\nu^2)-E^2(x_\nu)=\int_0^1 x_\nu^2 p(x_\nu)\ \mathrm{d}x_\nu-\left[\int_0^1 x_\nu p(x_\nu)\ \mathrm{d}x_\nu\right]^2$$

$$=\frac{1}{3}-\frac{1}{4}=\frac{1}{12}$$

[9] Clearly, $p(x_\nu|t_i)$ is a constant $i/N<x_\nu<(i+1)/N$ and zero otherwise. That this constant is N follows from the fact that the integral of $p(x_\nu|t_i)$ over all x_ν must be equal to 1.

Up until now we have been evaluating communication systems in terms of their signal-to-noise ratios S/N where, with the received waveform denoted by $y(t)=x(t)+n(t)$, the signal denoted by term $x(t)$, and the noise denoted by $n(t)$, the signal power S was defined as

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}x^2(t)\ \mathrm{d}t=<x^2(t)>$$

while the noise power N was determined from $<n^2(t)>$.

Actually, if the signal had had some nonzero average value c, we would have been interested, not in $<x^2(t)>$, but rather in $<[x(t)-c]^2> = <x^2(t)>-c^2$, since it is the variation about the average that conveys information, not the variation about zero. That is, suppose we arbitrarily added the constant c' to the signal. This would certainly not increase the amount of information at the receiver, but the signal power would be increased by an amount $<[x(t)+c']^2>-<x^2(t)>$. Thus, the definition of the signal-to-noise ratio that we have been using should properly be expressed as

$$\frac{S}{N}=\frac{<[x(t)-c]^2>}{<n^2(t)>}$$

where $c=<x(t)>$. The average value of the noise, of course, is zero, $<n(t)>=0$. But, since $<[x(t)-c]^2>=<x^2(t)>-[<x(t)>]^2=\sigma_x^2$ and $<n^2(t)>=\sigma_n^2$, we have

$$\frac{S}{N}=\frac{\sigma_x^2}{\sigma_n^2} \tag{4.51}$$

In the situation at hand, the term $n(t)$ at the output of the detector is less explicit. But clearly, if the desired signal is x_ν and the quantized signal $x_{\nu q}$ is received, then the "noise" is $x_\nu-x_{\nu q}$ and the noise variance is

$$E(x_\nu-x_{\nu q})^2-E^2(x_\nu-x_{\nu q})=\sigma_\nu^2 \tag{4.52}$$

(see footnote 10) while the signal variance remains

$$E(x_\nu^2)-E^2(x_\nu)=\sigma_s^2 \tag{4.53}$$

Thus the signal-to-noise ratio in this case becomes

[10] This, of course, assumes that $(A^2\Delta t)/(N_0/2)$ is large enough so that the probability of transmitting $x_{\nu q}$ and receiving $x_{\nu q}'$ where $x_{\nu q}'\neq x_{\nu q}$ is very small.

$$\left(\frac{S}{N}\right)_{\text{PDM}}=\frac{\sigma_s^2}{\sigma_\nu^2}=\left(\frac{T}{\Delta t}\right)^2=N^2 \tag{4.54}$$

so long as $\dfrac{A^2\Delta t}{\left(\dfrac{N_0}{2}\right)}$ is large.

It can be shown, by an argument identical to that used in the case of PAM, that the effective bandwidth per PDM channel is $B=1/2\Delta t$ cps. Since we are sending one sample of data every T seconds, the same information could be transmitted with a PAM (or SSB) bandwidth of $B_0=1/2T$ cps. The bandwidth expansion β using PDM is defined as the ratio of the amount of bandwidth necessary for PDM transmission to that which would be needed with PAM transmission. Consequently

$$\beta=\frac{B}{B_0}=\frac{T}{\Delta t} \tag{4.55}$$

and hence

$$\left(\frac{S}{N}\right)_{\text{PDM}}=\beta^2 \tag{4.56}$$

(approximately, so long as $A^2>5N_0B$). As with FM, the output signal-to-noise ratio increases in proportion to the expansion in bandwidth. Like FM, PDM also exhibits a threshold effect reflected in the fact that when the predetection signal-to-noise ratio $A^2/N_0B=(1/N)(A^2/N_0B_0)$ is too small, the probability of mistaking a positive pulse for a negative one, and vice versa, becomes large enough to degrade the performance significantly.

PULSE CODE MODULATION (PCM)

In the discussion of PDM it was found useful to quantize the data before transmission. Not only was the receiver much easier to build when the signal could have one of only a finite number of states, but when the signal-to-noise ratio became only moderately large the channel itself contributed negligible error, the primary error being only that of quantization. However, a method for obtaining the same performance at a considerably smaller expenditure of bandwidth suggests itself after the following consideration: Suppose the quantized PDM signal could have one of four states. Then, letting a 1 represent a positive pulse and a 0 a negative pulse, the four PDM signals are represented by the following configurations:

$$\begin{array}{c}1\ 0\ 0\ 0\\ 1\ 1\ 0\ 0\\ 1\ 1\ 1\ 0\\ 1\ 1\ 1\ 1\end{array}\tag{4.57}$$

But, just as it is more efficient to use a number system in which the position of a symbol as well as its form contributes to its value, it is useful to consider a communication system in which the same rules apply. In particular, if only two symbols are available, we can always represent the four symbols binarily as

$$\begin{array}{c}0\ 0\\ 0\ 1\\ 1\ 0\\ 1\ 1\end{array}\tag{4.58}$$

Each state now needs only one-half as many symbols to represent it. Clearly, if there are N states requiring N PDM binary symbols, only $\log_2 N$ (or the smallest integer equal to or larger than $\log_2 N$) PCM symbols are needed. (Of course, bases other than the binary could be employed, but binary systems possess certain distinct advantages and are by far the most commonly used.)

The analysis of this system parallels that of the PDM scheme. Suppose the signal is quantized into N levels where N is assumed to be a power of 2. Then, as before, when the signal-to-noise ratio is such that it is very unlikely that a positive pulse will be mistaken for a negative one or conversely, the primary contributor to the mean squared error at the receiver is the quantization error:

$$\sigma_n^2 = \frac{1}{12N^2}$$

where, again, the signal is assumed to vary between 0 and 1. Similarly, the signal variance remains

$$\sigma_s^2 = \frac{1}{12}$$

and the signal-to-noise ratio is

$$\left(\frac{S}{N}\right)_{\text{PDM}} = N^2 \tag{4.59}$$

As has been observed in the case of PDM and can be verified here, the effective bandwidth of a modulated signal consisting of sinusoidal pulses which can change amplitude only every Δt seconds is

$$B_{\text{eff}} = \frac{1}{2\Delta t} \tag{4.60}$$

But here $\Delta t = T/\log_2 N$ and, consequently, the bandwidth expansion factor is given by

$$\beta = \frac{B_{\text{eff}}}{B_0} = \frac{T}{\Delta t} = \log_2 N$$

Thus

$$\left(\frac{S}{N}\right)_{\text{PCM}} = N^2 = 2^{2\beta} \tag{4.61}$$

and the output signal-to-noise ratio increases not as β^2 but exponentially with β.

Again, as in PDM, this analysis is true only if the probability of mistaking a positive symbol for a negative symbol, and conversely is very small. In fact, this requirement is considerably more important in the case of PCM. While the signal symbolically represented by

1 1 1 1

might be received as

0 1 1 1

it still remains more like the transmitted signal than any of the other possible signals and no error is made. In contrast, if 1 1 is transmitted and 1 0 received in a PCM system, an error has been made. This is counteracted by the fact that the probability of making such an error is considerably less (for the same parameters) by using PCM than it is by using PDM because, for both systems, the probability of an error of this kind is

$$P_E = \frac{1}{2}\left\{1 - \operatorname{erf}\left[\left(\frac{A^2\Delta t}{N_0/2}\right)^{1/2}\right]\right\} \tag{4.62}$$

which is a monotonically decreasing function of Δt. But

$$(\Delta t)_{\text{PCM}} = \frac{T}{\log_2 N} = \frac{N}{\log_2 N}(\Delta t)_{\text{PDM}}$$

For large values of N, this advantage more than counteracts the above-mentioned disadvantage.

CODING AND WAVEFORM SELECTION

The last sections demonstrated the superior performance of quantized PDM and PCM systems. It was observed that the cost of an erronous reception of a pulse (or a symbol, as they are often denoted) can be high. Thus, the number of quantization levels, and hence the quantization error, is limited by the fact that Δt cannot be made too small or the symbol error probability becomes too great. Several techniques are available for decreasing the significance of symbol errors, and hence increasing the number of possible quantization levels. One of these is the use of error-correcting and error-detecting codes.

The subject of coding has received a large amount of attention in the literature in the past decade and is considerably too complex to be covered in any detail here. The principle of coding, however, is quite simple. Since one symbol error occurs relatively infrequently, two symbol errors will occur even more rarely, three still less often, etc. Thus, if instead of the set of symbols (eq. (4.58)), the symbols

$$\begin{array}{c} 0\ 0\ 0 \\ 0\ 1\ 1 \\ 1\ 0\ 1 \\ 1\ 1\ 0 \end{array} \tag{4.63}$$

were transmitted, it will be observed that one symbol error no longer causes an error in reception. If, for example, the sequence 0 0 1 is received, it is known that an error has occurred. It is not known which sequence was transmitted, but, at least, the error has been detected. Such a code is called an *error-detecting code.* If this code is altered further by adding two more symbols to each sequence in the following manner

$$\begin{array}{c} 0\ 0\ 0\ 0\ 0 \\ 0\ 1\ 1\ 0\ 1 \\ 1\ 0\ 1\ 1\ 0 \\ 1\ 1\ 0\ 1\ 1 \end{array} \tag{4.64}$$

it can be verified that, if only one symbol error is made, the resulting sequence is more like the original sequence than any other. If, for example, 0 0 0 0 0 is transmitted and one error is made so that 0 0 1 0 0 is received, it is seen that 0 0 1 0 0 differs from the first sequence in only one symbol, from the second in two, from the third in two, and from the fourth in five. Since one symbol error is made more likely than two or five, the most logical decision is that the transmitted sequence was 0 0 0 0 0. Then single errors can be corrected and codes which do this are called *single-error correcting codes.* These codes were achieved by adding *redundant* symbols to the original information symbols. By adding further redundancy, higher error-correcting properties can be obtained.

Note, however, that it is not necessarily true that this procedure does indeed decrease the probability of making a sequence error at the receiver. This is because, in order to keep the transmission rate the same, three and five symbols must be sent, using the codes (4.63) and (4.64), respectively, in the same time that two symbols would be sent using the code (4.58). Thus the time per symbol has dropped by a factor of 2/3 and 2/5, respectively, in the two cases. Since, from equation (4.62), the probability of a symbol error is a function of the symbol time, symbol errors are more likely using the codes (4.63) and (4.64). Thus, while error-detecting and error-correcting codes make symbol errors less costly, they also make them more probable. Whether there is a net gain or loss depends upon the code and upon the system parameters.

A second technique for decreasing the probability of an error at the receiver lies in the judicious selection of waveforms to represent the various signals. In the situation under consideration, the data source submits to the transmitter a signal which assumes one of N amplitudes and allows the transmitter T seconds to transmit it. It is desired to send this information as reliably as possible; i.e., to minimize the probability that the receiver interprets its input as something other than that which was transmitted. Intuitively, at least, the waveforms which represent the different amplitudes should look as different from each other as possible.

But what do we mean by "looking different from each other"? In "Spectra and Autocorrelation" in chapter 1, we introduced the concept of correlation as the measure of "likeness." Consider the *normalized cross-correlation coefficient* ρ_{12} between two time functions $y_1(t)$ and $y_2(t)$, where ρ_{12} is defined by (cf. "Expectation and Independence" in ch. 1):

$$\rho_{12}=\frac{\phi_{y_1y_2}(0)}{[\phi_{y_1}(0)\phi_{y_2}(0)]^{1/2}}$$

It may be shown that the two waveforms can be considered identical if their cross-correlation ρ_{12} is one. The smaller this normalized cross-correlation, the more dissimilar the waveforms are, until, when $\rho_{12}=-1$, the two waveforms are antipodal. Thus, one measure of similarity is the cross-correlation coefficient. It can, in fact, be shown that this is the measure of concern in the problem at hand.

So we have restated the problem as follows: We wish to find a mapping from the N possible signal amplitudes to N waveforms such that the cross-correlation between any two of these waveforms is as small as possible. The next three sections discuss some of the waveforms which may profitably be used for this purpose. Before proceeding, however, it is useful to obtain a standard whereby any N-level pulse-communication system may be judged.

Presumably, it is the maximum cross-correlation coefficient which will do the most damage, since this is the measure of the similarity of the two waveforms most likely to be mistaken for each other. Thus, we might ask, what is the minimum value this maximum normalized cross-correlation coefficient can obtain? Suppose we have N waveforms: $x_1(t), x_2(t), \ldots, x_N(t)$.

Then

$$\rho_{ij}=\frac{\int_0^T x_i(t)x_j(t)\,dt}{\left[\int_0^T x_i^2(t)\,dt\right]^{1/2}\left[\int_0^T x_j^2(t)\,dt\right]^{1/2}} \tag{4.65}$$

The average cross-correlation is clearly

$$\rho_{\text{ave}}=\frac{1}{N(N-1)}\sum_{\substack{i=1\\ i\neq j}}\sum_{j=1}\rho_{ij} \tag{4.66}$$

since there are $N(N-1)$ cross-correlation coefficients ρ_{ij}, where $i\neq j$. The ρ_{ii} terms are all equal to one and are excluded from the average because we are interested in the similarity between two different waveforms, not the trivial question of how similar a waveform is to itself. But

$$\rho_{\max}=\max_{\substack{ij\\ i\neq j}}\rho_{ij}\geq\rho_{\text{ave}}=\frac{1}{N(N-1)}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{ij}-N\right] \tag{4.67}$$

where the condition that $i\neq j$ has been replaced by subtracting the total contribution of the ρ_{ii} terms from the double summation. Substituting from equation (4.65) and interchanging the order of integration and summation, we find that

$$\frac{1}{N(N-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{ij}-N\right)=\frac{1}{N(N-1)}\left[\int_0^T\sum_{i=1}^{N}\left\{\frac{x_i(t)}{[\int_0^T x_i^2(t)\,\mathrm{d}t]^{1/2}}\right\}\right.$$
$$\left.\sum_{j=1}^{N}\left\{\frac{x_j(t)}{[\int_0^T x_j^2(t)\,\mathrm{d}t]^{1/2}}\right\}\mathrm{d}t-N\right] \quad (4.68)$$

The integral may then be written as

$$\int_0^T\left[\sum_{i=1}^{N}\left\{\frac{x_i(t)}{[\int_0^T x_i(t)\,\mathrm{d}t]^{1/2}}\right\}\right]^2\mathrm{d}t \quad (4.69)$$

and since the integrand is never negative, the integral is at least zero. Then

$$\rho_{\max}\geq\rho_{\text{ave}}\geq-\frac{1}{N-1} \quad (4.70)$$

We shall use this relationship as a standard for the evaluation of the sets of waveforms to be discussed in the succeeding sections.

FREQUENCY SHIFT KEYING (FSK)

Suppose, for the moment, that the signal x_ν is not quantized but again assumes all values $0<x_\nu<1$. A *frequency shift keyed* (FSK) communication system simply transmits a sinusoid with the frequency ax_ν to represent the signal x_ν. The received signal is then

$$y(t)=\sqrt{2}A\sin(\omega_c+ax_\nu)t+n(t) \qquad \nu T<t<(\nu+1)T \quad (4.71)$$

and the optimum receiver (cf. "Pulse Modulation Systems and Matched Filtering" in this chapter) involves the determination of the expression

$$\int_{\nu T}^{(\nu+1)T}\sqrt{2}y(t)\sin(\omega_c+ax_\nu)t\,\mathrm{d}t-\frac{A}{2}\int_{\nu T}^{(\nu+1)T}\sin^2(\omega_c+ax_\nu)t\,\mathrm{d}t \quad (4.72)$$

Differentiating this with respect to x_ν unfortunately does not establish an easily mechanized system, nor does it eliminate the necessity of determining the expression (4.72) explicitly for all values of x_ν. Therefore, for practical purposes, this technique too is restricted to quantized signals.

Suppose therefore that x_ν is equally likely to assume any of the N values, 0, $1/(N-1)$, $2/(N-1)$, . . . 1, and that

$$\omega_c=\frac{k_1\pi}{T}$$

and

$$a=\frac{(N-1)k_2\pi}{T}$$

so that $\omega_c+ax_\nu=k_3\pi/T$ for some integer k_3 regardless of the value of x_ν. Then consider the correlation between two signals $\sqrt{2}\sin(\omega_c+ax_i)t$ and $\sqrt{2}\sin(\omega_c+ax_j)t$ where $x_i=i/(N-1)$ and $x_j=j/(N-1)$:

$$\rho_{ij}=\frac{\int_0^T 2\sin(\omega_c+ax_i)t\sin(\omega_c+ax_j)t\,dt}{\left[\int_0^T \sqrt{2}\sin^2(\omega_c+ax_i)t\right]^{1/2}\left[\int_0^T \sqrt{2}\sin^2(\omega_c+ax_j)t\right]^{1/2}}$$

$$=\frac{1}{T}\int_0^T \cos a(x_i-x_j)t\,dt-\frac{1}{T}\int_0^T \cos(2\omega_c+ax_i+ax_j)t\,dt$$

$$=\begin{cases}0 & i\neq j\\ 1 & i=j\end{cases} \tag{4.73}$$

This set of waveforms, therefore, compares favorably with the optimum since for an optimum set

$$\max_{i\neq j}\rho_{ij}\geq-\frac{1}{N-1}\approx 0$$

for large N. (Other frequency displacements are, of course, possible, but neither the average nor the maximum cross-correlation can be significantly decreased.) Then the integrals

$$\int_{\nu T}^{(\nu+1)T}\sin^2(\omega_c+ax_\nu)t\,dt$$

are constant and independent of x_ν and the optimum decision is to select the largest over $x_i=0,\ 1/(N-1),\ 2/(N-1),\ \ldots\ 1$, of the integrals

$$Z_i=\int_{\nu T}^{(\nu+1)T} y(t)\sqrt{2}\sin(\omega_c+ax_i)t\,dt \tag{4.74}$$

Since $y(t)=\sqrt{2}A\sin(\omega_c+ax_\nu)t+n(t)$

$$E(Z_i)=\int_{\nu T}^{(\nu+1)T}2A\sin(\omega_c+ax_\nu)t\sin(\omega_c+ax_i)t\,dt$$

$$+\sqrt{2}\int_{\nu T}^{(\nu+1)T}E[n(t)]\sin(\omega_c+ax_i)t\,dt$$

$$\equiv\mu_i=AT\rho_{i\nu}=\begin{cases}0 & x_i\neq x_\nu\\ AT & x_i=x_\nu\end{cases}\tag{4.75}$$

and

$$\sigma_i^2=E(Z_i^2)-E^2(Z_i)=\frac{N_0}{2}T\equiv\sigma_0^2\tag{4.76}$$

Again, since $n(t)$ is a Gaussian random process, the variables Z_i are Gaussianly distributed. It can be shown that the probability of a correct decision at the receiver is just the product of the probabilities that all the variables Z_i, $i\neq\nu$, are less than the variable Z_ν. For a fixed value of Z_ν, this is

$$Pr(Z_i<Z_\nu|i\neq\nu,\ Z_\nu)$$

$$=\int_{-\infty}^{Z_\nu}p(Z_0)\,dZ_0\int_{-\infty}^{Z_\nu}p(Z_1)\,dZ_1\ldots\int_{-\infty}^{Z_\nu}p(Z_{N-1})\,dZ_{N-1}\tag{4.77}$$

where

$$p(Z_i)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(Z_i-\mu_i)^2}{2\sigma_i^2}\right]=\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{Z_i^2}{2\sigma_0^2}\right)\tag{4.78}$$

and

$$Pr[Z_i<Z_\nu|\ i\neq\nu]=\int_{-\infty}^{\infty}Pr[Z_i<Z_\nu|\ i\neq\nu,\ Z_\nu]p(Z_\nu)\,dZ_\nu$$

$$=\int_{-\infty}^{\infty}\left[\int_{-\infty}^{Z_\nu}p(Z_0)\,dZ_0\right]^{N-1}p(Z_\nu)\,dZ_\nu$$

$$=\frac{1}{(2\pi)^{N/2}\sigma_0^N}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{Z_\nu}\exp\left(-\frac{Z_0^2}{2\sigma_0^2}\right)dZ_0\right]^{N-1}$$

$$\exp\left[\frac{(Z_\nu-AT)^2}{2\sigma_0^2}\right]dZ_\nu$$

$$=\frac{1}{(2\pi)^{N/2}}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{W_\nu+[A^2T/(N_0/2)]^{1/2}}\exp\left(-\frac{W_0^2}{2}\right)dW_0\right]^{N-1}$$

$$\exp\left(\frac{W_\nu^2}{2}\right)dW_\nu\equiv P_N\left(\frac{A^2T}{N_0/2}\right)\tag{4.79}$$

where the substitutions $W_0 = Z_0/\sigma_0$ and $W_\nu = (Z_\nu - AT)/\sigma_0$ have been made. Clearly the probability of a correct decision is a function of N and $A^2T/(N_0/2)$ only.

The effective bandwidth occupancy of orthogonal [11] FSK is obviously $N/2T$, since from equation (4.73) all frequencies separated by an amount $1/2T$ can be used without mutual interference and N, such frequencies are needed for an N-level system. Thus the bandwidth occupancy of FSK and PDM are the same. The probability of an error is considerably greater with the latter system, however.

A slight variation on the modulation scheme described above is to use not only the signals $x_i(t) = \sqrt{2}A \sin(\omega_c + ax_i)t$, but also their negatives, $-x_i(t) = -\sqrt{2}A \sin(\omega_c + ax_\nu)t$. (See footnote 12.) Clearly, this does not increase the bandwidth requirement, yet it does double the number of available signals. The detector remains unchanged, except that the decision now involves the selection of the largest of the quantities $|Z_i|$ where Z_i is defined as in equation (4.72). The probability of a correct decision changes slightly and is given by

$$
\begin{aligned}
P_C &= Pr[|Z_i| < Z_\nu | i \neq \nu] \\
&= \int_0^\infty \left[\int_{-Z_\nu}^{Z_\nu} p(Z_0)\, dZ_0 \right]^{(N/2)-1} p(Z_\nu)\, dZ_\nu \\
&= \frac{1}{(2\pi)^{N/2}} \int_0^\infty \left[\int_{-W+[A^2T/(N_0/2)]^{1/2}}^{W+[A^2T/(N_0/2)]^{1/2}} \exp\left\{-\frac{W_0^2}{2}\right\} dW_0 \right]^{(N/2)-1} \\
&\quad \exp\left(-\frac{W^2}{2}\right) dW \\
&\equiv Q_N\left(\frac{A^2T}{N_0/2}\right) \qquad (4.80)
\end{aligned}
$$

Note that here in order to have N signals, we need only $N/2$ values of x_i, since for each x_i there are two possible waveforms. Therefore, for the same number of signals, biorthogonal codes require only one-half the bandwidth of orthogonal codes.

Equations (4.79) and (4.80) will be further investigated later in this chapter.

[11] Waveforms satisfying the property that their cross-correlation coefficients are all identically zero are called *orthogonal*, and the ensemble of waveforms is called an *orthogonal code.*

[12] This set of waveforms can easily be shown to have the property that

$$
\rho_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \text{ and } x_i(t) \neq -x_j(t) \\ -1 & x_i(t) = -x_j(t) \end{cases}
$$

The ensemble of these waveforms is called a *biorthogonal code.* Note that $\rho_{\text{ave}} = -1/(M-1)$, where M is the number of waveforms, and hence meets the bound (4.70).

PULSE POSITION MODULATION (PPM)

Another modulation system which could, in principle, be used to transmit continuous data, but for practical purposes is limited to quantized data, is pulse position modulation. This modulation scheme consists of transmitting a pulsed sinusoid of duration T/N seconds, the position of which conveys the information. In the quantized case only N nonoverlapping pulse positions are allowed. The envelopes of the pulses are shown in figure 4.9 for the case in which $N=4$.

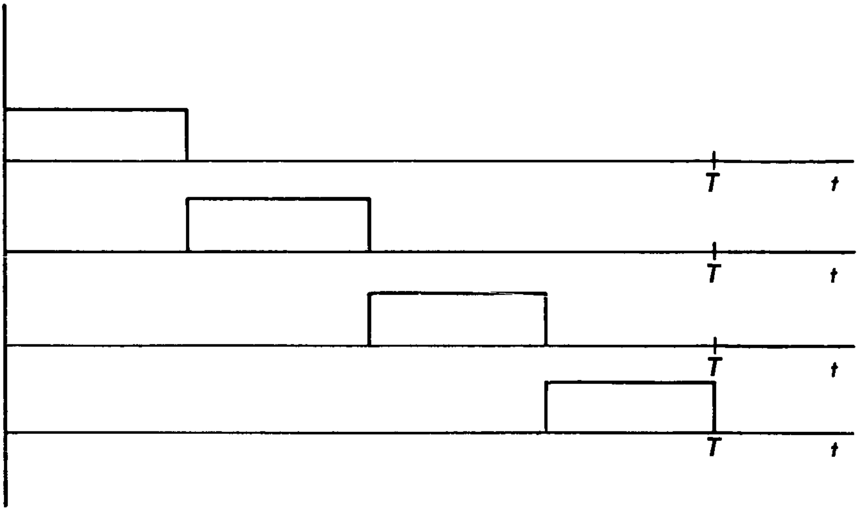


FIGURE 4.9—PPM four-level waveforms.

It will be observed that since no two different waveforms are nonzero at the same instant of time, the cross-correlation coefficient ρ_{ij} is identically zero for all $i\neq j$ (see eq. (4.65)). Since the waveforms may be denoted

$$\sqrt{2}A' \sin \omega_c t \begin{cases} \dfrac{iT}{N}<t<(i+1)\dfrac{T}{N} \\ 0 \quad \text{Otherwise} \end{cases}$$

for each $i=0, 1, \ldots, N-1$, it is seen that each represents an average power of $(A')^2T/N$ (where it is again assumed that $\omega_c=k\pi N/T$ for some integer k). Thus, letting $(A')^2=A^2N$, we have exactly the same situation as that encountered in the investigation of FSK and the same conclusions apply. The bandwidth occupancy is again $N/2T$ and the error probability expression (4.79) still applies. The difference between the two systems, of course, lies in their instrumentation and in the fact that the peak power requirements of PPM is N times as great as that for FSK if they are to exhibit the same average power and hence the same performance.

FURTHER EXCHANGE OF BANDWIDTH FOR RELIABILITY

The orthogonal codes of the last two sections have demonstrated a method for decreasing the probability of an error at the receiver. Since N-level PCM requires a bandwidth of $\log_2 N/2T$ cps while the orthogonal signal used an $N/2T$ cps bandwidth, it is seen that the increase in reliability is at the cost of an $N/\log_2 N$ increase in bandwidth.

As the astute reader may have already observed, however, there is no imperative reason why the number of levels of quantization N must be equal to the number of waveforms to be used. That is, suppose the data are quantized and each sample is represented as a sequence of $\log_2 N$ binary digits, as in PCM. (It is convenient, although not necessary, to limit the discussion to binary digits.) Then, so far as the transmitter is concerned, the input is a sequence of binary digits. The fact that they may be grouped into blocks of $\log_2 N$ digits to represent the sample does not necessarily affect the way in which they should be transmitted. Each m consecutive digit may be transmitted as one of $M=2^m$ waveforms (there are 2^m different sequences of m binary digits), where M may be greater or less than N. The receiver decides which of the M waveforms was transmitted, reconstructs the corresponding m digits and thereby presents the original sequence at the output. The user of the information then proceeds to divide the received sequence into the blocks of n digits corresponding to the data. It is often quite inefficient to require that $M=N$ since N may vary from one data source to another. If M were always equal to N, the communication system might have to be different for each source. It will be seen in the next chapter, in fact, that one of the advantages of digital operation is that the telemetry system can be made relatively independent of the data.

If, then, N is fixed, it is of interest to determine the effect of increasing M. To determine this, we observe that, once both the sampling rate T and N, and hence $n=\log_2 N$, have been determined, the amount of time available for transmitting each binary digit (or bit) of information is $T/\log_2 N=T_b$. If m bits are to be transmitted as one waveform, each such waveform must last exactly $mT_b=(\log_2 M)T_b$ seconds. The bandwidth occupancy for M signal orthogonal codes is

$$\frac{M}{2T_W}=\frac{M}{\log_2 M}\left(\frac{1}{2T_b}\right)=\frac{M\log_2 N}{N\log_2 M}\frac{1}{2T_N}$$

where T_W is the time available per waveform, and T_N is the time available per waveform when $M=N$. Thus the bandwidth expansion is proportional to $M/\log_2 M$ and the ratio of the bandwidth required for an arbitrary m bit grouping to that necessary if each data signal were to be transmitted separately with an orthogonal code is $M \log_2 N/N \log_2 M$.

To see the advantages of using larger values of M, the error probability $P_E=[1-P_M(A^2T_b/N_0)]$ is plotted in figure 4.10 for various values of $m=\log_2 M$ as a function of A^2T_b/N_0, rather than $A^2T/(N_0/2)$ as in equation (4.79), since the time per bit T_b is the quantity fixed by the data requirements rather than the time per waveform. It is seen that it is always advantageous to increase the value of M. Figure 4.11 presents the analogous results for biorthogonal codes; $P_E'=[1-Q_M(A^2T_b/N_0)]$. It is seen that the latter codes are always superior to the former, but that this advantage is insignificant for $M>3$. Note that for the case $M=2$, P_E' is just the error probability given by equations (4.47) and (4.48) for PDM and PCM, since there, too, there were just two antipodal signals. In the PDM case, T_b becomes $\Delta t=T/M$; for PCM, $T_b=T/\log_2 M$.

SYNCHRONIZATION

It will be noted that all of the systems analyzed have depended upon a common time reference (i.e., synchronization) for the transmitter and receiver. In particular, it was assumed that it is possible to generate a sinusoid at the receiver which has the exact phase and frequency of the received carrier. This information is usually called *carrier* synchronization. Further, optimum demodulation depends on the knowledge of the instants of time in which a waveform begins and ends (i.e., the instants of time νT, $\nu=0, 1, \ldots$). This is called *word* or *waveform* synchronization. Other timing references are often necessary, too. If, for example, the waveform conveys m bits of information where each sample is quantized to $N=2^n$ levels, as in the previous section, the data user must have available synchronization information which will enable him to separate the received sequence of bits into the right blocks of n bits. In addition, if, as is almost always the case, numerous data sources are to be multiplexed together, the user must be able to identify which part of the data corresponds to which source. This is referred to as *frame* synchronization.

Numerous synchronization techniques have been explored. One of the most common and straightforward methods involves the use of a phase-locked loop to track an unmodulated sinusoid, or other periodic signal, the phase and frequency of which convey the necessary timing. Further discussion of the synchronization problems is presented in chapter 6 in which the related problem of ranging is investigated.

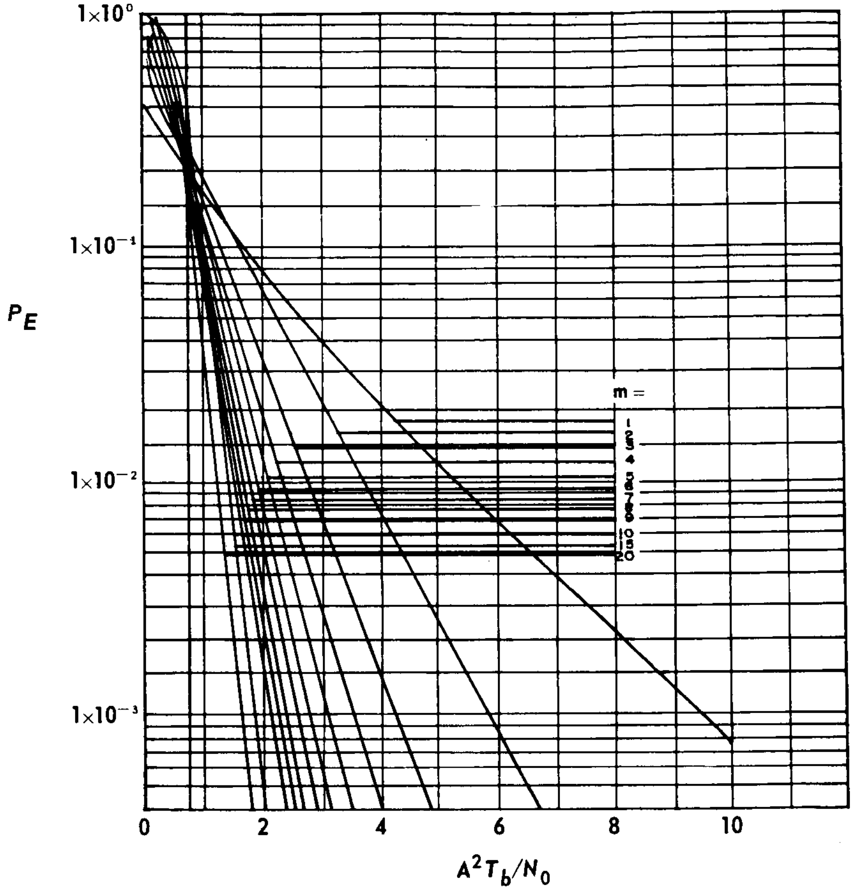


FIGURE 4.10—Orthogonal codes: word error probability.

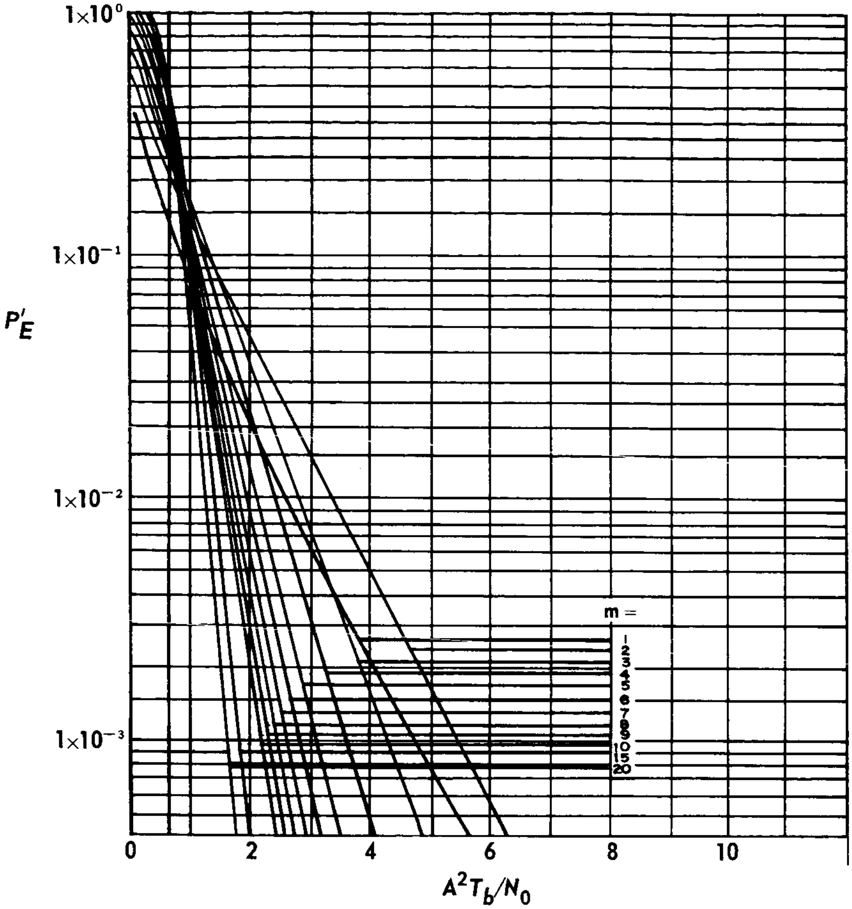


FIGURE 4.11—Biorthogonal codes: word error probability.

CHAPTER 5

Data Compression

ALL OF OUR EFFORTS in investigating methods for improving the efficiency of long-distance communication have been concerned with the communication system itself. But there is another area in which great improvements can theoretically be made; viz, in the preprocessing of the data which are to be communicated. The sampling theorem states that the data must be sampled at least every $T=1/2W$ seconds. Unfortunately, W is not generally known for the data sources being monitored. Indeed, it is precisely the uncertainty and the difficulty in predicting the behavior of the sources that make them worthwhile to observe. Thus, W, and therefore the sample rate T, must be estimated, and to insure that important information is not lost, they must be conservatively estimated with the result that a source is usually sampled considerably more rapidly than it would ideally be. In addition, even if W were known and the sampling rate were optimally determined, successive samples would, in general, be rather highly dependent. That is, much of the information contained in the $(i+1)$th sample would have been predictable from the ith, $(i-1)$th, $(i-2)$th samples. If these samples themselves are transmitted, much effort is wasted in sending *redundant* information. The effects of redundancy are readily observed in written text: CNSDR FR XMPL THS SNTNC. Few will find any difficulty in reading that array of letters even though more than 30 percent of the symbols have been deleted. That much greater elimination of redundancy is possible is evidenced by the various shorthand and speedwriting techniques which have been developed. Any telemetry system that transmitted conventionally written text would be sending much redundant information and thereby using its power inefficiently.

The suggestion then is that shorthand techniques should be used aboard a spacecraft so only essential information is transmitted. Since fewer symbols would be sent, more time could be spent per symbol, thereby decreasing the probability of an error at the receiver and hence increasing the reliability of the system.

This argument for the removal of redundancy seems at variance with the discussion of the previous chapter in which it was stated that, at least under certain conditions, redundant symbols could be advantageously added to reduce the likelihood of an error at the receiver. The difference in the two situations is in the fact that the added redundancy can be carefully selected for the maximum improvement, whereas the naturally occurring redundancy, while providing some error-correcting ability, is far from efficient. Suppose, for example that the following sequence of temperatures were received: 250.031°, 250.033°, 250.035°, 350.033°, 250.034° . . . . The error in the most significant position of the fourth reading is easily corrected, yet an error in the least significant position would probably pass unnoticed. As an alternative method, suppose only the first reading were sent followed by the incremental temperature changes, each repeated three times as follows: 250.031°, +2 +2 +2, +2 +2 +2, −2 −2 +2, +1 +2 +1, ... The same number of symbols are involved in both cases (actually, the sign is one of only 2 symbols, rather than 1 of 10), yet the error in the least significant position of the fourth and fifth readings is readily discernible. In the latter case, the natural redundancy was replaced with controlled redundancy, thereby increasing the correctability of small errors. This system is not necessarily practical. Changes of more than 0.009° cannot be indicated and in this format, if an error does occur, all succeeding readings will be in error, since the incremental changes are transmitted rather than the readings themselves. Nevertheless, it hopefully does illustrate the point. This chapter will discuss several techniques for the elimination of redundancy or, as it is variously called, *data compression* or *data compaction.*

Nth DIFFERENCES

In the example of the preceding section, the incremental changes, or *first differences,* of the data samples were transmitted rather than the samples themselves. This method may be extended by transmitting the second differences (i.e., the first differences of the first differences) or third differences, etc. The advantage of this procedure, of course, depends upon the data itself. If, for example, the data are increasing, or decreasing, nearly linearly $x_\nu \approx a+bt$, then the first differences are nearly constant, the second differences are nearly all zero. Transmission of the second differences in this case would generally require no more than one symbol. In general, if the data are varying approximately as an $(n-1)$th polynomial in time, $x_\nu = a_1 + a_2t + \ldots + a_nt^{n-1}$, the nth differences are nearly constant and may be transmitted by a very few symbols. It will be observed that in order to reconstruct the original data from the nth differences, the first n data samples (or their equivalent) must also be transmitted. For example, if third differences are transmitted, it is seen by referring to table 5.1 that all the data samples can be reconstructed if either the underlined or the bracketed information is transmitted.

TABLE 5.1—A Third Difference Data Compression Scheme

Samples	1st differences	2d differences	3d differences
$\underline{[x_1]}$			
	$[x_1-x_2]$		
$\underline{x_2}$		$[x_1-2x_2+x_3]$	
	x_2-x_3		$\underline{[x_1-3x_2+3x_3-x_4]}$
$\underline{x_3}$		$x_2-2x_3+x_4$	
	x_3-x_4		$\underline{[x_2-3x_3+3x_4-x_5]}$
x_4		$x_3-2x_4+x_5$	
	x_4-x_5		$\underline{[x_3-3x_4+3x_5-x_6]}$
x_5		$x_4-2x_5+x_6$	
	x_5-x_6		
x_6			

The disadvantages of the *n*th differences scheme are apparent. First, because of the tendency of errors to propagate as observed earlier, the process must be periodically truncated and begun again. But a more serious problem is that it is seldom possible to predict that the data will closely approximate an *n*th order polynomial for any particular value of *n*. Because the data frequently have a constant nonzero mean, and because the data must generally be sampled too often in order to insure that no information is lost, the first-difference schemes are sometimes practical, but seldom are higher order difference schemes used.

RUN-LENGTH ENCODING

It has been observed that, because of the lack of knowledge concerning the bandwidth of the data sources, the data must generally be sampled too often. The redundancy can be considerably reduced by taking the first differences of the samples. Even so, however, because the data bandwidth is difficult to predict, it is not known in advance how different successive samples will be. Thus, it is not known how many symbols should be allowed for each first difference; that is, will two successive samples most likely differ only in the least significant position or the two least significant positions, or will they sometimes differ in all positions? Presumably, all of these situations will occur at various times. But if the first differences demand nearly as many symbols as did the original samples, there is little advantage in taking first differences.

One method for partially overcoming this difficulty is by the technique of *run-length encoding.* Suppose the data samples are quantized and represented by binary digits. The first differences are also binary digits, but since presumably two successive samples do not usually differ significantly, the first differences will consist mainly of zeros, particularly in the most significant positions. The run-length encoding scheme. takes advantage of this fact by transmitting, not the first differences themselves but rather the spacings between successive "ones" in the binary sequence representing these first differences. Thus suppose the data samples are 0 1 0 1 1 1 0 1, 0 1 0 1 1 1 0 0, 0 1 0 1 1 0 1 1, 0 1 0 1 1 0 1 1, 0 1 0 1 1 1 0 0 . . ., where the first digit represents the sign of the sample. The first differences are then . . . 0 0 0 0 0 0 0 1, 0 0 0 0 0 0 0 1, 0 0 0 0 0 0 0 0, 1 0 0 0 0 0 1, . . . and the run-length code, obtained by counting the spacing between successive ones in the first-difference sequences and representing the spacings in binary form, is 1 1 1, 1 1 1, 1 0 0 0, 1 1 0, ... The number of digits has thereby been reduced from 32 to 13 without sacrificing the ability to transmit an 8-digit first difference if necessary. This example is somewhat misleading, since we have ignored the necessity of providing the commas in the run-length encoded sequence. One not particularly efficient method for overcoming this difficulty is to keep the encoded subsequence length constant. That is, suppose we divide the code into groups of three digits and constrain it so that no run of zeros greater than six can be represented by one set of three digits. If the binary number 7 is transmitted, the sequence immediately following it is to be added to it to determine the distance between successive ones in the original sequence. Thus, if the distance between successive ones is 6, the sequence 1 1 0 is transmitted; 7 is represented by 1 1 1, 0 0 0; 8 by 1 1 1, 0 0 1; 13 by 1 1 1, 1 1 0; 14 by 1 1 1, 1 1 1, 0 0 0, etc. The data sequence example used previously is now encoded as 1 1 1 0 0 0 1 1 1 0 0 0 1 1 1 0 0 1 1 1 0, requiring 21 symbols. No special information is needed here to decode this sequence, however, other than the knowledge of its beginning, since it is always divided into groups of three. It will be noted that this scheme is not constrained to be used only in conjunction with the first-differencing technique. It is applicable whenever long runs of zeros (or 1's) occur frequently in the information to be transmitted.

HUFFMAN CODES

Another method for reducing the average numbers of digits to be transmitted without sacrificing the ability to transmit all possible messages lies in the use of *Huffman codes.* Observe in the example of the last section that the most commonly occurring first-difference sequences are apparently 0 0 0 0 0 0 0 0, 0 0 0 0 0 0 0 1, and 1 0 0 0 0 0 0 1. In probable descending order of occurrences, the possible remaining sequences are: 0 0 0 0 0 0 0 1 0, 1 0 0 0 0 0 0 1 0, 0 0 0 0 0 0 1 1, 1 0 0 0 0 0 1 1, 0 0 0 0 0 1 0 0, 1 0 0 0 0 1 0 0, etc. The rationale for the Huffman encoding process is as follows: The average number of symbols necessary to represent the data will be reduced by representing the most common sequences by fewer symbols than the less-common sequences. This is best illustrated by an example. Suppose the sequences

0 0

0 1

1 0

1 1

occur with the probabilities 1/2, 1/4, 1/8, and 1/8, respectively. If the sequences themselves are transmitted, the average number of digits per sequence is, of course, two. But suppose the following identification is made:

$$\begin{aligned} 0\ 0 &\rightarrow 1 \\ 0\ 1 &\rightarrow 0\ 1 \\ 1\ 0 &\rightarrow 0\ 0\ 1 \\ 1\ 1 &\rightarrow 0\ 0\ 0 \end{aligned} \tag{5.1}$$

Then the average number of symbols transmitted is

$$1\,Pr(0\ 0) + 2\,Pr(0\ 1) + 3\,Pr(1\ 0) + 3\,Pr(1\ 1) = 1.75$$

not a vast improvement perhaps, but one which becomes considerably more impressive as the sequence length increases and the difference between the probability of the most probable sequence and that of the least probable increases. It may be wondered why the particular identification of the expression (5.1) was chosen rather than some other. Why, in particular, were not more of the two-symbol sequences used? The reason lies in the fact that any random ordering of the symbol sequences on the right side of the expression (5.1) can be uniquely deciphered. This follows from the observation that each symbol either ends in a "1" or contains three zeros. Three consecutive zeros are therefore recognized as corresponding to the sequence 1 1, while the end of any other symbol will be identified by a 1. This is not true for all possible mappings of the form (5.1). The Huffman coding algorithm, however, guarantees that such a mapping can always be accomplished. The optimum mapping depends upon the probabilities of occurrence of the original sequences. Nevertheless, even if these probabilities are not known, some saving can generally be effected by mapping those first-difference sequences corresponding to the least change between successive samples onto the shortest symbol sequences, those corresponding to greater differences onto longer sequences. This method, like the one of the previous section, clearly is not restricted to be used only on the first differences of the data. It is useful whenever the different sequences which are to be transmitted occur with different probabilities.

THE FLOATING BARRIER TECHNIQUE

Another method for data compression which we shall consider briefly in this far-from-exhaustive discussion is called the *floating barrier* scheme. With this method a quantization level is set and an initial sample is transmitted. The data are periodically sampled, as before, but successive samples are transmitted only if they differ from the last transmitted sample by more than the predetermined quantization. The operation of the process is illustrated in figure 5.1. The sampling interval is Δt and the quantization amplitude is q. The transmitted samples are circled.

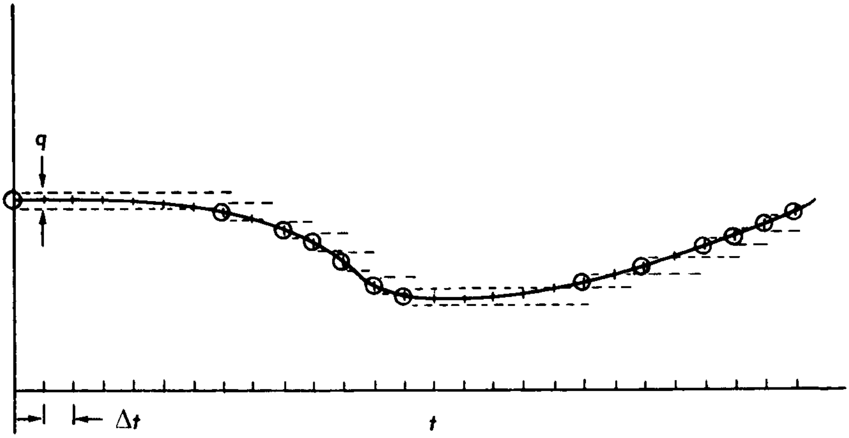


FIGURE 5.1—The floating barrier scheme.

It is seen that for this particular example the number of samples which must be transmitted is reduced from 27 to 13, yet the signal is always known to be within $\pm q/2$ of the sample points.

It is not sufficient here just to send the samples. Since every sample is not transmitted, it is also necessary to provide the number of samples which were omitted between successive transmitted samples if it is to be possible to reconstruct the original signal. This, of course, somewhat decreases the savings in the number of digits which must be transmitted.

TELEVISION COMPRESSION

Before leaving the subject of data compression, it is well to consider one case in which some rather impressive savings can be made; namely, in the transmission of video information. It has been verified that the intensity of the light representing a television picture can be quantized to only eight levels from black to white with little picture degradation. If there are 32 quantization levels, the eye cannot see the difference between the quantized picture and one in which a continuum of intensity levels is allowed. A commercial television picture is quantized spatially into 512 horizontal lines. If it were further quantized vertically, also into 512 lines for example, a picture could be represented in digital form as an array of 512×512 numbers which, assuming eight-level intensity quantization, could vary from 0 to 7, or in binary form, from 0 0 0 to 1 1 1. Thus it would take $512\times512\times3=786\,432$ binary digits to represent one television picture. But, clearly, these digits are not independent. A black spot is likely to be surrounded by black spots, a white spot by white spots. Furthermore, on the relatively rare occasions in which two adjacent spots do differ, they most generally differ by only one intensity level. This dependence exists, of course, not only spatially but temporally if successive pictures of the same object are transmitted, and not only between adjacent symbols but over a considerably greater distance. Clearly, then, there is room for a vast reduction in the amount of data which actually must be transmitted. By what factor the data may be compressed and by what techniques it can be accomplished depends to a large degree on the assumptions made about what is being observed. Reductions by a factor of a hundred or more in the number of bits necessary to represent a picture have been calculated for some special situations.

DESTRUCTIVE COMPACTION

The methods of data compression we have considered up to now are sometimes referred to as *nondestructive* data compression techniques. All the information from the source is to be transmitted; only redundancy is to be removed. Often, however, the user is not interested in all the available information. Perhaps it is only the maximum and minimum temperatures of some device, for example, that are of interest. If this is the case, it is certainly wasteful to send the complete set of temperature readings. It is relatively easy to determine the extremes on board the spacecraft and to send those readings only. Similarly, if it is desired to know only the statistical distribution of the counts from a Geiger counter, it is certainly not efficient to transmit a complete record of the number of counts per second. Rather, a data-reducing mechanism aboard the vehicle should be used to process the information to determine the desired histogram which can then be transmitted.

This sort of data reduction is called *destructive* data compression, since not all of the information provided by the source is used. The percentage of savings can be extremely high for this kind of data compression for cases in which the user needs only a quite specific subset of the information which could be provided. As an extreme case, suppose it were necessary to know only the average Geiger-counter reading over a period of 10 hours. The amount of transmitted data could be reduced by a factor of 36 000 by sending only this average rather than sending, for example, a reading every second.

The limitation of this procedure, however, lies in the fact that the user seldom knows what aspect of the data will be useful to him before he sees it. He is generally reluctant to allow any information to be destroyed for fear of missing some important but unsuspected observation.

SUMMARY AND CONCLUSIONS

Numerous methods have been investigated for the elimination of redundancy in data which are to be telemetered. Among them are included:

(1) The *n*th difference method whereby the data samples themselves are not transmitted, but rather their first, second, or higher order differences.

(2) The process of run-length encoding in which the presumed predominance of zeros in the data (or in their first or higher order differences) is used to reduce the number of symbols transmitted.

(3) Huffman encoding which relies upon the generally valid assumption that some data samples (or commonly, data-sample first differences) occur more frequently than others.

(4) The floating barrier technique which transmits samples only when they are significantly different from the last transmitted sample.

(5) Destructive compression techniques.

The amount of reduction in the transmitted data afforded by compaction varies widely from source to source. Compaction ratios as high as a thousand or more are possible with nondestructive methods; destructive compression ratios are almost unlimited depending upon the amount of information actually needed.

Each of the methods discussed involves several difficulties. One of the more serious is the tendency of errors to propagate. In order that (nondestructive) data compression be significant, a considerable amount of statistical information concerning the source[13] must be available (information which is generally not entirely available when applied to spacecraft sources, since each experiment encounters rather unpredictable situations). This, coupled with the fact that data compaction equipment considerably complicates the spacecraft electronics,[14] has in the past limited data compaction efforts. Nevertheless, as data-handling capabilities of spacecraft are to be significantly increased, data compression techniques will become more and more attractive.

[13] This fact has given considerable impetus to the investigation of self-adaptive data compression schemes which have the ability to vary in accordance with the source statistics which are estimated on board the spacecraft.

[14] All of the above techniques, for example, are more effective when the data are changing slowly than when the data are rapidly changing. Thus the data rates will vary depending upon the sources. Since the transmitter must generally operate at a fixed rate, buffering or temporary storage facilities must be provided.

CHAPTER 6

Spacecraft Tracking

THE PROBLEM OF TRACKING or keeping a running record of the position and velocity of a spacecraft is not generally classified as a telemetry problem. The word "telemetry" usually implies the transmission of information through space from one position to another. Yet, spacecraft tracking involves the same equipment and many of the same techniques as does telemetry; it logically falls under the study of telemetry systems.

Tracking can be accomplished with one antenna or several. If two or more antennas are involved, it is theoretically possible to obtain all information concerning the position of the vehicle by pointing all the antennas at it and observing their respective orientations. The antennas may "see" the vehicle either from its radio transmission or by its reflection of radar signals from the ground. This technique suffers from two disadvantages. First, it does require two or more antennas suitably placed, with a means of communicating to a common point the orientations of each antenna, in order to determine the position of the object being observed. And second, it is difficult to obtain the desired accuracy when the spacecraft is fairly distant from the earth. Even a 0.01° error in the orientation of each of the antennas can result in a sizable error in the estimate of the position of the spacecraft as shown in figure 6.1. Since a successful mission strongly depends upon the ability to place the spacecraft on the desired trajectory, which in turn depends upon the precision with which its position can be measured, more precise methods must be considered.

Since, as seen in figure 6.1, the distance from the earth, in particular, is subject to large ambiguities when it is determined from triangulation, it is apparent that another means of determining this distance or range would be useful. Such a technique, long used in conjunction with conventional radar, is to transmit a narrow pulse toward the vehicle and measure the time which elapses before the echo returns. Since these time delays can be measured electronically to a very high precision, the distance can be determined quite accurately. The ultimate limitation in measuring distance by this method lies in the precision with which the velocity of electromagnetic radiation is known.

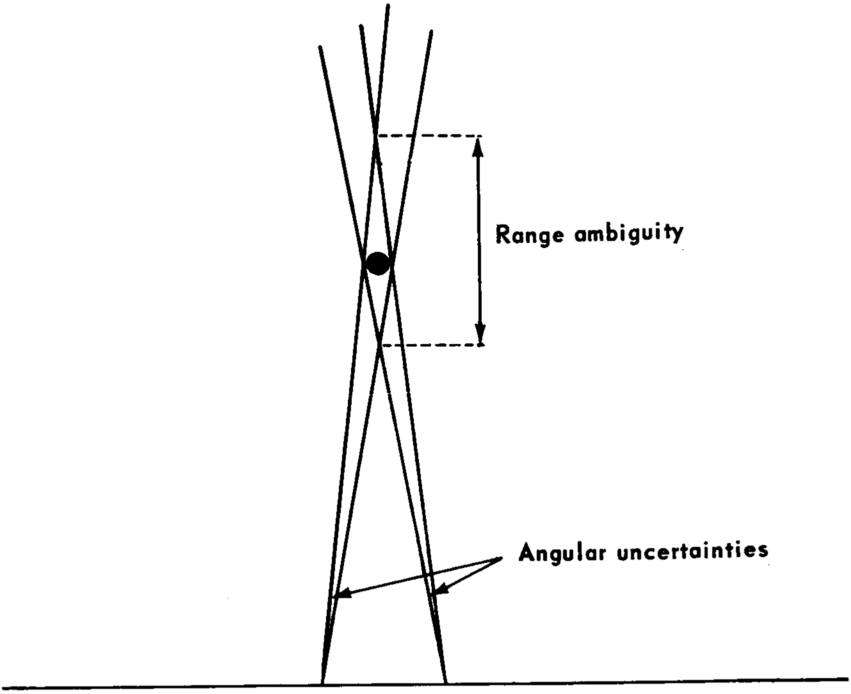


FIGURE 6.1—Range ambiguity in triangulation.

Several practical improvements can be made on this technique of ranging. In order to understand these improvements, it is necessary to consider more precisely how ranging information is extracted from a reflected pulse train.

PULSED RADAR

In chapter 4 we considered, in some detail, the structure of matched filter detectors. It was argued that the optimum detector for estimating which of the signals $x_i(t) \quad (i=1, 2, \ldots)$ was most likely transmitted when the received signal was $y(t)$ involved the formulation of the integrals

$$I_i = \int y(t) x_i(t) \, dt \tag{6.1}$$

and the selection of the largest of these.

The determination of range involves essentially the same considerations. Here, assuming the transmitted pulse is rectangular

$$x(t) = \begin{cases} \sqrt{2}B \sin \omega_c t & 0 < t < \Delta t \\ 0 & \text{Otherwise} \end{cases} \tag{6.2}$$

the received signal is one of the continuum of signals

$$y_s(t)=\begin{cases}\sqrt{2}A\ \sin\,\omega_c(t-t_0) & t_0<t<t_0+\Delta t\\ 0 & \text{Otherwise}\end{cases}\tag{6.3}$$

where t_0 can have any value over the range of ambiguity, $t_1<t_0<t_2$. Thus letting $y(t)=y_s(t)+n(t)$ represent the received signal plus noise, the optimum detector determines the integrals

$$I(\hat{t}_0)=\int_{\hat{t}_0}^{\hat{t}_0+\Delta t} y(t)\sqrt{2}\ \sin\,\omega_c(t-\hat{t}_0)\ dt\tag{6.4}$$

for all values of $\hat{t}_0$ and selects the largest. We note, again, that since $n(t)$ is a Gaussian process with $E[n(t)]=0$ and $E[n(t)n(t+\tau)]=(N_0/2)\delta(\tau)$, $I(\hat{t}_0)$ is a Gaussianly distributed random variable with

$$\begin{aligned}E[I(\hat{t}_0)]&=2A\int_{t_0}^{\hat{t}_0+\Delta t}\sin\,\omega_c(t-t_0)\ \sin\,\omega_c(t-\hat{t}_0)\ dt\\ &=A[\Delta t-(t_0-\hat{t}_0)]\ \cos\,\omega_c(t_0-\hat{t}_0)\\ &\quad-\frac{A}{2\omega_c}\left\{\sin\,(2\omega_c\Delta t+\omega_c\hat{t}_0-\omega_c t_0)-\sin\,(\omega_c\hat{t}_0-\omega_c t_0)\right\}\\ &=A[\Delta t-(t_0-\hat{t}_0)]\ \cos\,\omega_c(t_0-\hat{t}_0)\end{aligned}\tag{6.5}$$

where it is assumed that $\hat{t}_0<t_0<\hat{t}_0+\Delta t$ and that $\omega_c=\pi/\Delta t$. If $\hat{t}_0-\Delta t<t_0<\hat{t}_0$

$$E\{I(\hat{t}_0)\}=A[\Delta t-(\hat{t}_0-t_0)]\ \cos\,\omega_c(\hat{t}_0-t_0)$$

and hence, in general

$$E\{I(\hat{t}_0)\}=\begin{cases}A(\Delta t-|t_0-\hat{t}_0|)\ \cos\,\omega_c(t_0-\hat{t}_0) & \hat{t}_0-\Delta t<t_0<\hat{t}_0+\Delta t\\ 0 & \text{Otherwise}\end{cases}\tag{6.6}$$

In addition, it is easily verified that

$$\begin{aligned}\sigma_n{}^2&=E[I^2(\hat{t}_0)]-E^2[I(\hat{t}_0)]\\ &=N_0/2\Delta t\end{aligned}\tag{6.7}$$

The expected value of the integrator output is illustrated in figure 6.2 as a function of the time delay estimate $\hat{t}_0$.

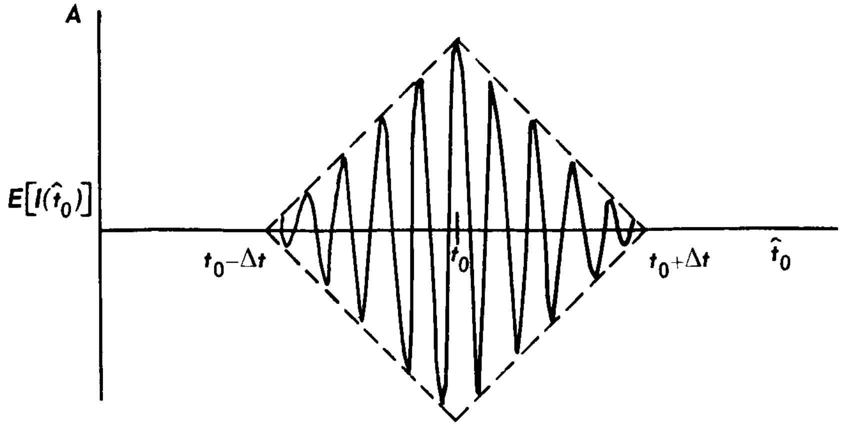


FIGURE 6.2—Pulse-matched-filter output.

The case in which $\hat{t}_0 \approx t_0$ should be distinguishable, with a prescribed reliability, from all other competitors. This, in turn, puts a certain requirement upon the signal-to-noise ratio $A^2\Delta t/(N_0/2)$. To be more specific, suppose that the requirements are to be able to distinguish correctly between the cases in which $\hat{t}_0 = t_0$ and $\hat{t}_0 = t_0 + \Delta t$ with some specified small error probability. Since $I(t_0)$ and $I(t_0+\Delta t)$ are both Gaussian random variables with

$$E[I(t_0)] = A\Delta t$$

$$E[I(t_0+\Delta t)] = 0$$

and

$$E[I^2(t_0)] - E^2[I(t_0)] = E[I^2(t_0+\Delta t)] - E^2[I(t_0+\Delta t)]$$
$$= N_0/2\Delta t$$

this is identical to the situation in which a two-word orthogonal code set is used in a communication system ("Frequency Shift Keying" and "Further Exchange of Bandwidth for Reliability" in ch. 4). That is, the probability of mistaking $I(t_0+\Delta t)$ for $I(t_0)$ is just that of transmitting one of two orthogonal code words and making an error in identifying it at the receiver, assuming, of course, that the signal-to-noise ratio is the same in the two cases. Thus, if $S/N = 2(A^2\Delta t/N_0) = 16$, the probability of confusing $I(t_0)$ with $I(t_0+\Delta t)$ is approximately 2.4×10^{-3} (fig. 4.10).

The primary limitation to this procedure is in the amount of peak power the transmitter is able to radiate. This limitation can be partially overcome by sending the pulses periodically with a repetitive rate T which is greater than the range of ambiguity, $T > t_2 - t_1$. Then, the summation

$$\sum_{i=0}^{N-1} I_i(t_0) = \sum_{i=0}^{N-1} \int_{\hat{t}_0+iT}^{\hat{t}_0+\Delta t+iT} y(t)\sqrt{2}\,\sin\,\omega_c(t-\hat{t}_0)\,\mathrm{d}t \tag{6.8}$$

may be performed, increasing the effective integration time, and, hence, the signal-to-noise ratio by a factor of N. In order to do this, the vehicle being ranged must either be relatively stationary, or the relative motion between the receiver and the vehicle during the interval between pulses must be known and compensated for. This latter may be accomplished by transmitting, in addition to the pulse, some unmodulated sinusoid of amplitude $\sqrt{2}a$ at the frequency ω_c. This sinusoid can be tracked, for example, with a phase-locked loop and the loop output used to control the clock generating the pulses as shown in figure 6.3. Note that the tracked sinusoid is also used to eliminate the high-frequency structure from the pulse (cf. fig. 6.2).

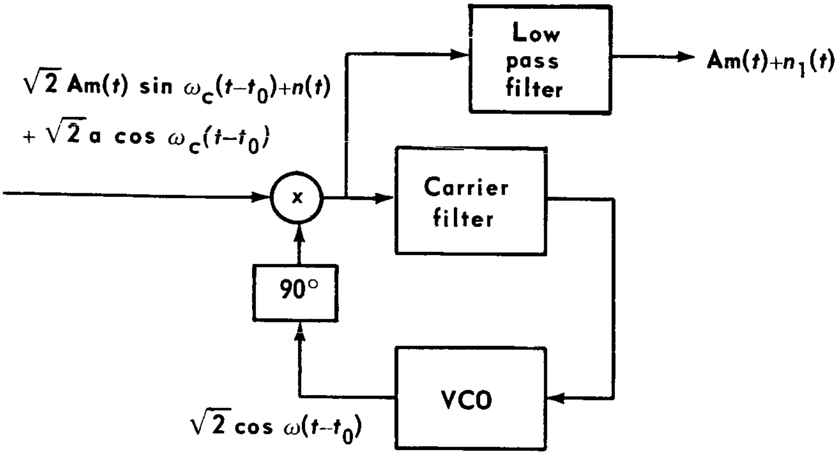

$\sqrt{2}\,Am(t)\,\sin\,\omega_c(t-t_0)+n(t)$
$+\sqrt{2}\,a\,\cos\,\omega_c(t-t_0)$
x
Low pass filter
$Am(t)+n_1(t)$
Carrier filter
90°
VCO
$\sqrt{2}\,\cos\,\omega(t-t_0)$

FIGURE 6.3—Ranging signal demodulator.

This is done primarily for convenience of instrumentation, since the integrals of equation (6.4) now become

$$I'(\hat{t}_0) = \int_{\hat{t}_0}^{\hat{t}_0+T} \hat{m}(t)y'(t)\,\mathrm{d}t = \int_{\hat{t}_0}^{\hat{t}_0+\Delta t} y'(t)\,\mathrm{d}t \tag{6.9}$$

where

$$y'(t) = Am(t)+n_1(t),$$

$$\hat{m}(t) = \begin{cases} 1 & \hat{t}_0<t<\hat{t}_0+\Delta t \\ 0 & \text{Otherwise} \end{cases}$$

$$m(t) = \begin{cases} 1 & t_0<t<t_0+\Delta t \\ 0 & \text{Otherwise} \end{cases}$$

The continuum of integrals can now be replaced by one followed by a sampler

$$I'(\hat{t}_0) = \int^{\hat{t}_0+\Delta t} y'(t)\ \mathrm{d}t - \int^{\hat{t}_0} y'(t)\ \mathrm{d}t$$
$$= Y(\hat{t}_0+\Delta t) - Y(\hat{t}_0) \tag{6.10}$$

Thus, only the single integration

$$Y(t') = \int^{t'} y'(t)\ \mathrm{d}t \tag{6.11}$$

need be determined. The variance of $I'(t_0)$ remains the same as that of $I(t_0)$, but the mean becomes

$$E[I'(\hat{t}_0)] = \begin{cases} A(\Delta t - |t_0 - \hat{t}_0|) & t_0 - \Delta t < \hat{t}_0 < t_0 + \Delta t \\ 0 & \text{Otherwise} \end{cases} \tag{6.12}$$

and is thus the envelope of the expected value of the function $I(\hat{t}_0)$. This removal of the fine structure does not decrease the resolution, since this information is still contained in the knowledge of the phase of the signal sin $\omega_c(t-t_0)$. That is, the locally generated signal cos $\omega_c(t-t_0)$ can be used to control the clock gating the sampler which samples the output of the integrator. If the carrier phase and the pulse phase are coherent, it is only necessary to sample at the instants of time $\hat{t}_0 = 2\pi i/\omega_c$, where $i = 1, 2, \ldots$, since a pulse cannot begin at any other instant of time. Since the phase of the carrier in general varies quite slowly, the bandwidth of the loop tracking the signal $\sqrt{2a}$ cos $\omega_c(t-t_0)$ can be quite small and, consequently, a is usually small compared with A, the amplitude of the received pulse.

The primary disadvantage to this ranging procedure as observed is that all the signal power must be sent in a relatively short period of time. If the repetition rate is T and the pulse width is Δt, pulse energy is being transmitted only $(\Delta t/T)\times 100$ percent of the time. The peak power radiated must be increased by a factor of $T/\Delta t$ if the average is to be kept the same as that which would be possible with continuous radiation.

But how can continuous radiation be used to achieve the desired range resolution? One answer lies in the use of *pseudo-random sequences.*

PSEUDO-RANDOM SEQUENCES

To begin, we observe that the autocorrelation function of any wave train, consisting of pulses whose amplitude while varying from pulse to pulse remains fixed for the constant pulse duration Δt, exhibits the property that

$$\phi(\tau)=\phi(n\Delta t)\left[\frac{(n+1)\Delta t-\tau}{\Delta t}\right] \\ +\phi[(n+1)\Delta t]\left[\frac{\tau-n\Delta t}{\Delta t}\right] \qquad n\Delta t\leq\tau\leq(n+1)\Delta t \qquad (6.13)$$

We have already observed this behavior for the pulse train discussed in "Spectra and Autocorrelation" in chapter 1. Note that it also applies to the wave train considered in the previous section. The details of the proof of this statement (6.13) are left as an exercise for the reader.

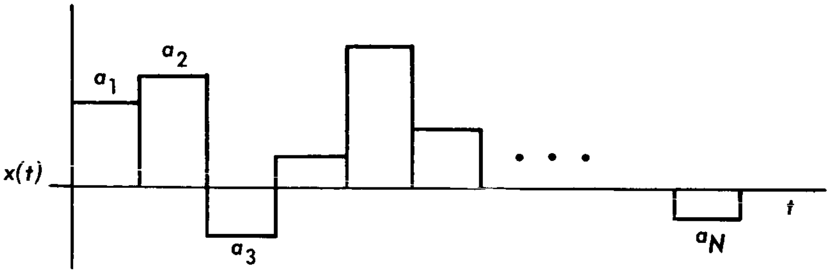


FIGURE 6.4—A sequence of fixed amplitude pulses.

This observation allows us to consider the autocorrelation function only at the points $n\Delta t$, $n=0$, 1, . . ., since we know that $\phi(\tau)$ varies linearly between these points and that $\phi(\tau)=\phi(-\tau)$. In addition, a pulse train of the type considered here may be represented simply by the amplitude of the successive pulses. Thus, for example, the wave train in figure 6.4 can be represented by the sequence $a_1\ a_2\ a_3\ .\ .\ .\ a_N$, where a_i denotes the amplitude of the ith pulse. If the pulse train is periodic with period T, as will be assumed here, $a_{N+j}=a_j$ where $N=T/\Delta t$ is the number of pulses per period. Similarly, the autocorrelation function $\phi(i\Delta t)$ is given by

$$\phi(i\Delta t)=\int_0^T x(t)x(t+i\Delta t)\ dt \\ =\Delta t\sum_{j=1}^{N} a_j a_{j+i} \qquad (6.14)$$

Then the autocorrelation function, too, can be readily determined from the knowledge of the amplitudes of the pulses which comprise the wave train in question.

Consider now the periodic pulse train represented by the sequence 1, 1, −1 as shown in figure 6.5(a). The autocorrelation function is easily determined

$$\phi(0)=\Delta t\sum_{j=1}^{3}a_j^2=3\Delta t$$

$$\phi(\Delta t)=\Delta t\sum_{j=1}^{3}a_ja_{j+1}=-\Delta t$$

$$\phi(2\Delta t)=\Delta t\sum_{j=1}^{3}a_ja_{j+2}=-\Delta t$$

and repeats with period $T=3\Delta t$. The autocorrelation function is illustrated in figure 6.5(b).

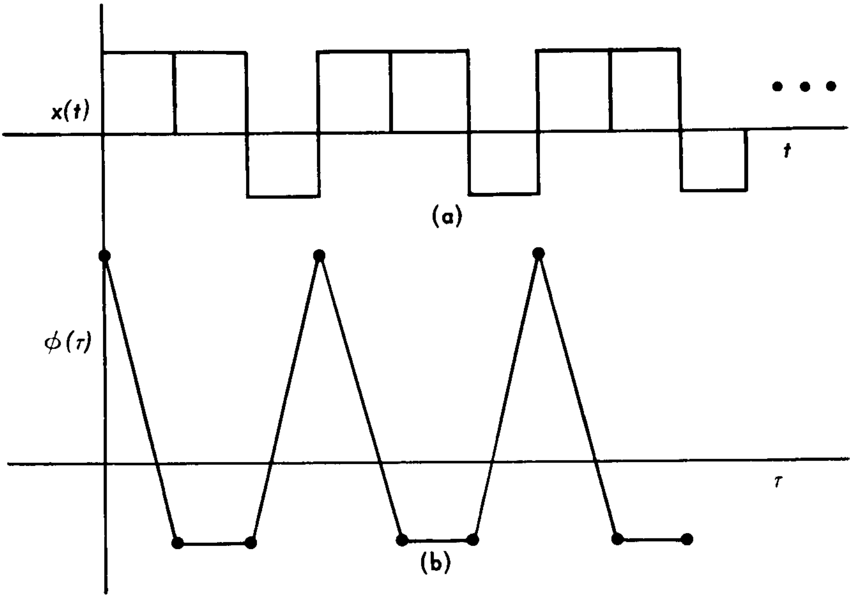


FIGURE 6.5—(a) The periodic pulse train $x(t)$; (b) its autocorrelation function $\phi(\tau)$.

The reader may verify that the sequence

$$1\ \ 1\ \ 1\ \ -1\ \ -1\ \ 1\ \ -1$$

when repeated periodically has the autocorrelation function

$$\phi(0)=7\Delta t$$

$$\phi(i\Delta t)=-\Delta t \qquad i=1,\ 2,\ 3,\ 4,\ 5,\ 6$$

It is, in fact, possible to construct sequences whose periodic autocorrelation function has the values

$$\begin{aligned} &\phi(0) = N\Delta t \\ &\phi(i\Delta t) = -\Delta t \qquad i = 1, 2, \ldots, N-1 \end{aligned} \tag{6.15}$$

for all values of $N = 2^n - 1$, where $n = 1, 2, \ldots$. The sequences are called *pseudo-random* or PR sequences because of the many properties which they have in common with truly random two-level sequences.

The reader may have observed that any one of these sequences, combined with all its periodic phase shifts, comprises a set of N waveforms whose maximum normalized cross-correlation is $\phi(i\Delta t)/\phi(0) = -1/N$, thereby nearly attaining the bound on the maximum cross-correlation derived in the section entitled "Coding and Waveform Selection" in chapter 4. An additional property of these sequences is that they contain $(N+1)/2$ pulses with the amplitude $+1$ and $(N-1)/2$ pulses with the amplitude -1. Thus the cross-correlation between any phase shift of one of these sequences and the sequence consisting of only -1's is

$$-\Delta t \sum_{i=1}^{N} a_i = -\Delta t$$

Consequently, the set of all periodic phase shifts of such a sequence plus the sequence consisting of all -1's contains $N+1$ waveforms with the maximum normalized cross-correlation equal to $-1/N$ and, hence, actually does attain the bound derived in "Coding and Waveform Selection" in chapter 4. These sequences can be used to generate waveforms in the manner described there. To illustrate, consider the set of sequences

$$\begin{array}{ccc} -1 & -1 & -1 \\ 1 & 1 & -1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \end{array}$$

which represent the waveforms shown in figure 6.6.

Before returning to the ranging problem and the application of PR sequences to it, let us observe another quite interesting feature of these sequences: their ease of generation. The diagram of figure 6.7 represents a set of three storage cells which retain a number, zero or one, until a clock pulse causes a shift. When the shift or clock pulse occurs, each binary digit shifts to the right, and the modulo-two sum of the contents

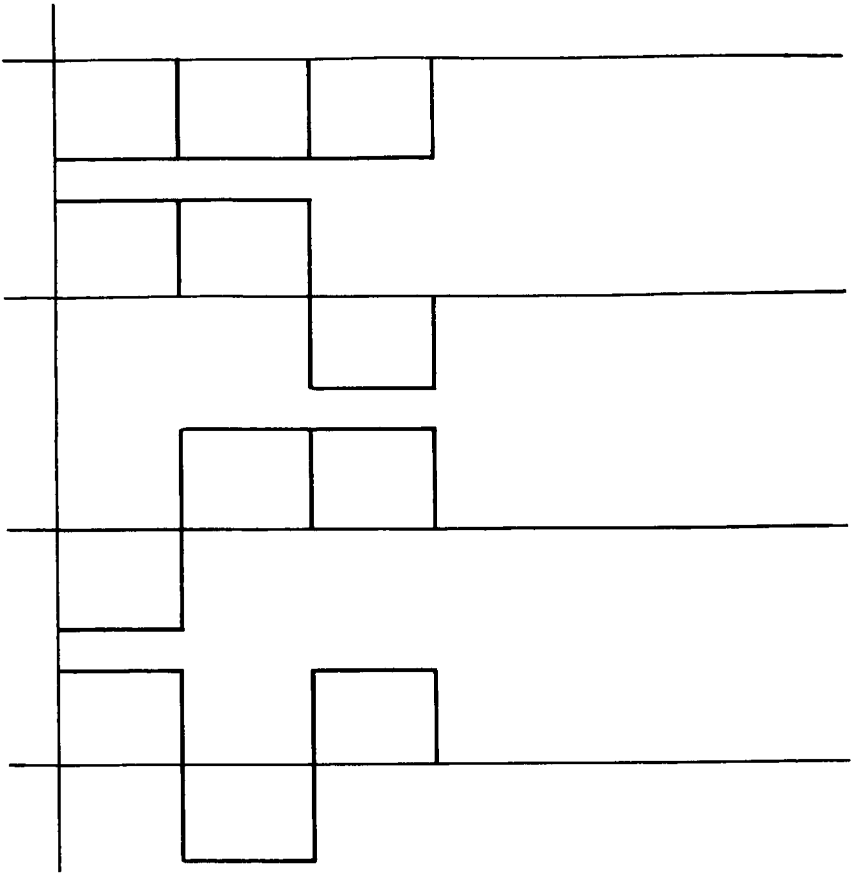


FIGURE 6.6—A set of waveforms achieving the bound in "Coding and Waveform Selection," chapter 4

of the last two cells is shifted into the first. By modulo-two sum, we mean that

$$
\begin{array}{cccc}
0 & \oplus & 0 & =0 \\
0 & \oplus & 1 & =1 \\
1 & \oplus & 0 & =1 \\
1 & \oplus & 1 & =0
\end{array}
$$

The reader may verify that the mechanism of figure 6.7 called a *shift register* does indeed generate the complete PR sequence given earlier of length seven regardless of which three binary digits are originally placed in the cells, unless all three digits are zero. (To be consistent with the earlier notation, the output "0" of this device must be converted to "-1.") In general, a sequence of length $N=2^n-1$ may be generated by an n-cell shift register.

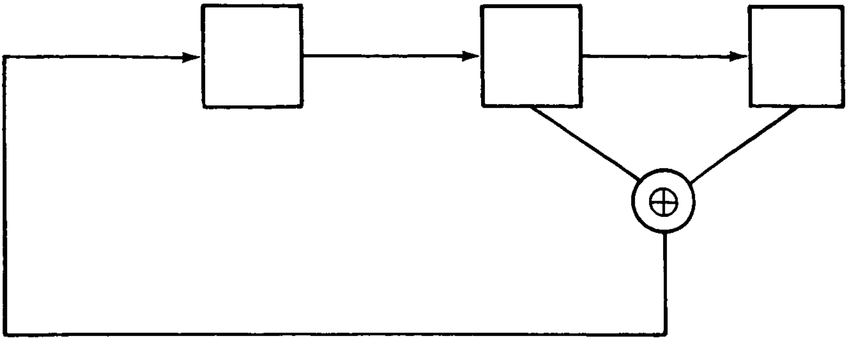


FIGURE 6.7—A three-stage shift register.

RANGING WITH PSEUDO-RANDOM SEQUENCES

The application of pseudo-random sequences to ranging should be fairly obvious from the discussion of the previous section. Since the periodic autocorrelation of PR sequences assumes the values

$$\phi(0)=N\Delta t$$

$$\phi(i\Delta t)=-\Delta t \qquad i=1, 2, \ldots, N-1$$

The complete autocorrelation function is shown in figure 6.8.

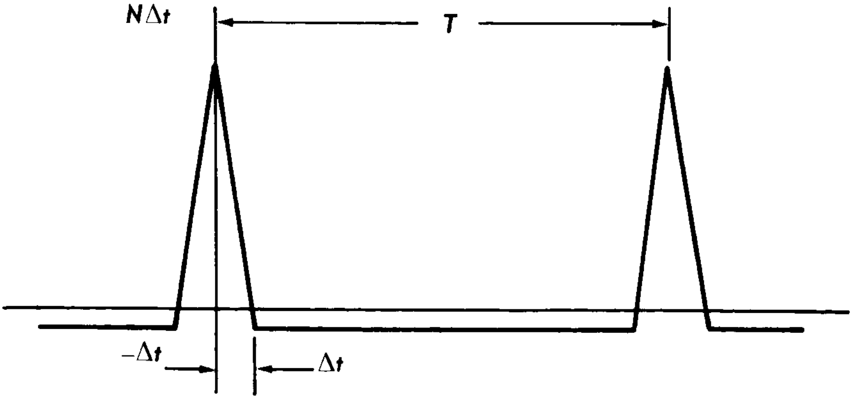


FIGURE 6.8—The PR-sequence autocorrelation function.

Suppose the $\hat{m}(t)$ in equation (6.9) were this sequence rather than a single pulse of duration Δt. Then, since the optimum detector forms the integral

$$I'(\hat{t}_0)=\int_{\hat{t}_0}^{\hat{t}_0+T} \hat{m}(t)y'(t)\,\mathrm{d}t=\int_{\hat{t}_0}^{\hat{t}_0+T} x(t-\hat{t}_0)y'(t)\,\mathrm{d}t \qquad (6.16)$$

where $y'(t)=A'x(t-t_0)+n_1(t)$ and $x(t)$ represents the PR sequence, it follows that

$$\begin{aligned} E[I'(\hat{t}_0)] &= A'\int_{\hat{t}_0}^{\hat{t}_0+T} x(t-t_0)x(t-\hat{t}_0)\,\mathrm{d}t=A'\phi(t_0-\hat{t}_0) \\ &= NA'\left[\Delta t-\left(\frac{N+1}{N}\right)|t_0-\hat{t}_0|\right] \qquad t_0-\Delta t<\hat{t}_0<t_0+\Delta t \\ &= -A'\Delta t \qquad \text{Otherwise} \qquad (6.17) \end{aligned}$$

Note the similarity between this equation and equation (6.12). If the average power radiated in the two cases is equated, then $NA'=A$; and if N is large, as it must be for high-range resolution, then $A'=A/N$ is negligible compared with A and the two cases are essentially equivalent. The PR-sequence technique has the definite advantage that the same amount of power is always being radiated; the peak power is equal to the average power.

In ranging space probes some distance from the earth, it is difficult to reflect from them the amount of power necessary for high-resolution ranging. However, the spacecraft can be equipped with a *transponder* which receives the transmitted signal, tracks both the carrier and the PR sequence, and regenerates them for retransmission back to earth. This technique increases the amount of power received at the earth tracking station by several orders of magnitude and is, in fact, the technique most commonly used in conjunction with planetary missions.

It is well to observe at this point the close relationship between ranging and synchronization. The latter necessitates the establishment of a common time reference between the transmitter and the receiver, while ranging involves the measurement of the signal transit time between the two. Both, therefore, rely upon the determination of the frequency and phase of a received signal. The frequency determination enables the receiver clock to be controlled so that it runs at the same rate as the transmitter clock; the phase measurement, obtained, for example, by identifying the phase of a received PR sequence, establishes a common point in time. The PR sequence used in the ranging case must be long enough to resolve the distance ambiguity. That is, if the sequence length corresponds to only 10 miles distance, it may be difficult to resolve the ambiguity as to whether the vehicle being ranged is, for example, 200 010 miles away or 200 020 miles away. If, on the other hand, its length corresponds to 100 000 miles, there would probably be little difficulty in determining whether the vehicle is 200 010 or 300 010 miles away. Similarly, in the synchronization problem, the sequence length must be great enough to resolve all time ambiguities not easily resolved otherwise.

DOPPLER MEASUREMENTS

If the transmitter, transmitting signals with a carrier frequency of f_c cps, and the receiver are moving away from each other with a relative velocity of v meters/sec, then in time T the distance between them has increased by an amount vT while f_cT carrier cycles have been transmitted. Designating the carrier *wavelength* by $\lambda = c/f_c$ where c is the velocity of light, we see that while f_cT cycles are transmitted, the distance has increased by vT/λ cycles and, hence, only $f_cT - vT/\lambda$ cycles are received. Thus, the apparent carrier frequency at the receiver is

$$f_c\left(1-\frac{v}{c}\right)$$

The *Doppler frequency* Δf, the difference between the transmitted frequency and the received frequency, is therefore

$$\Delta f = \frac{v}{c} f_c \tag{6.18}$$

If the transmitter and the receiver are moving *toward* each other with a velocity v, the Doppler shift is, of course, the same in absolute magnitude but represents an apparent increase rather than a decrease in the received frequency.

In the case of the ranging of a space vehicle, in which the transmitted signal is either reflected back to the earth or received and retransmitted to the earth by the vehicle, the Doppler shift has been effected twice and hence yields a shift of $2v/c$ cps at the earthbound receiver. By measuring the received frequency, then, it is possible to measure the velocity of the spacecraft toward or away from the earth. This can be most easily accomplished by forming the product of the transmitted signal carrier $\sqrt{2} \sin \omega_c t$ with the received signal carrier $\sqrt{2}A \sin [(\omega_c \pm 2\pi\Delta f)t + \phi]$ obtaining

$$A \cos (2\pi\Delta f t \pm \phi) - A \cos [(2\omega_c \pm 2\pi\Delta f)t + \phi] \tag{6.19}$$

After filtering out the double-frequency component, the frequency of the remaining term $A \cos (2\pi\Delta ft \pm \phi)$ can be determined by passing it through a bank of very narrow bandpass filters

$$H_i(j2\pi f) = 1 \qquad f_0 + i\delta f < f < f_0 + (i+1)\delta f$$
$$i = 0, 1, \ldots, N-1 \qquad (6.20)$$

where f_0 is the lowest and $f_0 + N\delta f$ is the highest expected Doppler shift. The filter which exhibits the greatest average output power then indicates the value of Δf and, thereby, the component of the velocity of the vehicle toward or away from the earth. If the sign of the frequency shift is in doubt, it can also be determined by some additional manipulation.

Since the carrier frequency is quite high, often on the order of 1000 Mc/sec, determining the Doppler frequency to the nearest cps, for example, establishes the velocity of the vehicle to within less than 1 mph. Doppler measurements, when combined with some initial information concerning the vehicle position and velocity at some point in time and with the knowledge of the trajectory which the vehicle must follow, provide a complete record of the position of the vehicle. The limitation to this approach in tracking a space vehicle lies in its dependence upon the need for rather precise measurements of the initial position and velocity data, upon the knowledge of the gravitational fields to which the vehicle is subject, and upon the mathematical difficulties in using all of this information to calculate the trajectories. Direct range measure-

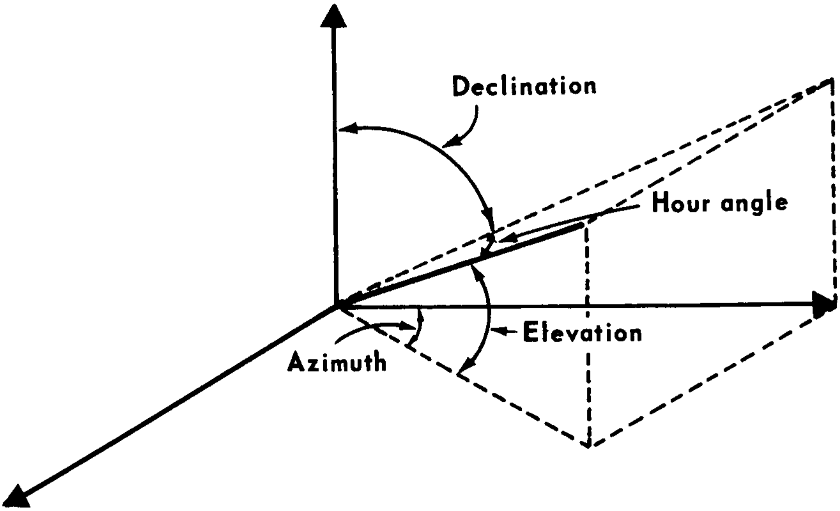


FIGURE 6.9—Two antenna coordinate systems.

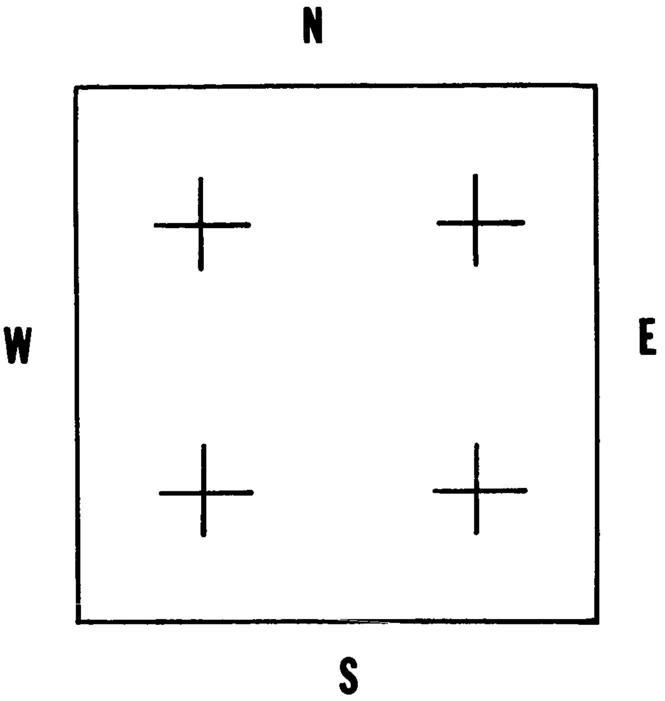


FIGURE 6.10—Antenna pickup.

ments as described in the previous sections are also subject to the latter two limitations when it is desired to determine the trajectory of the vehicle. The Doppler measurements and the direct range measurements do provide complementary information, however, and the combination of these two measurements can often be used to improve the estimates of the physical parameters involved in the calculation of these trajectories.

ANGLE MEASUREMENT

As mentioned earlier, part of the knowledge of the position of the vehicle may depend upon the ability to measure the angular position of the antenna pointed directly at it; at any rate, it is necessary to be able to keep the antenna accurately directed at the vehicle in order to realize its high gain. The coordinate system in which the antennas move depends, of course, upon the type of mount used to support them. The two most common coordinate systems, the azimuth-elevation (az-el) and the hour angle-declination (HA–DEC) systems are illustrated in figure 6.9. The solid line represents the direction in which the antenna is pointing. The planes of motion of the antenna can be visualized by keeping one of the coordinates fixed and moving the other of the same pair.

Since parabolic antennas are highly directional, the coordinates can be varied until the received signal power is a maximum. While it is possible to point the antenna fairly precisely at the vehicle in this manner, it is difficult to achieve the accuracy desired. Moreover, since the vehicle is generally moving in the antenna coordinate system, it is desirable to provide a method whereby the antenna can automatically be kept aimed at the source. It is, therefore, useful to have some indication from the received signal itself about the error in the antenna orientation, an indication which would enable this error to be decreased as well as to aid in tracking the moving vehicle.

This indication can be provided quite simply by placing at the antenna focus point not one, but four pickup points oriented as shown in figure 6.10. The signals from these four points are combined in three ways: (1) the difference between the sum of the outputs of the two "north" and that of the two "south" pickups, (2) the difference between the sum of the outputs of the two "east" and that of the two "west" pickups are formed, as well as (3) the sum of the outputs of all four pickups. If the antenna is properly oriented, the signals (1) and (2) will be zero.

If the antenna is pointed too far north, more signal power will be reflected toward the north pickups than to the south, thereby generating an error signal (1). This signal can be used to indicate the directional error causing the antenna to be shifted south until there is no more error signal. The same comments, of course, apply to orientation errors in the other directions.

CHAPTER 7

Pioneer, Mariner, and the DSIF

IN THIS CONCLUDING CHAPTER we shall discuss the communication systems which have actually been used in the Pioneer and Mariner programs, as well as the ground-based equipment known as the DSIF.

THE PIONEER COMMUNICATIONS SYSTEM

In this section we briefly investigate the telecommunication system used on the Pioneer IV lunar probe. Since this system is now obsolete we shall not go into great detail, but simply outline the fundamental concepts involved.

The Pioneer IV communications system consisted of three FM channels which were frequency multiplexed, the combined signal being used to phase-modulate a carrier at 960 Mc/sec.

Although there were only three channels, a total of seven different data sources were monitored through the use of time sharing and by a process of superimposing several signals simultaneously on one subcarrier and using the information redundancies to separate the data at the receiver. The first channel consisted of a temperature measurement. Because it was known that the temperature would not undergo any rapid changes, several "event indicators" were superimposed upon this signal. The events in question caused a jump in the temperature amplitude. These jumps could not be due to the temperature; they were recognized as the desired event indications and were subtracted out, leaving only the temperature measurements. The several events to be designated in this way were: (1) an indication of the initiation of a "despinning" operation in which weights were extended to slow down the spinning of the probe; (2) a step voltage change indicating that the optical devices had received light due to the closeness to the moon; and (3) a series of steps (positive and negative), indicating that the optical cells were periodically seeing light and then darkness in accordance with the spin rate of the vehicle as it passed by the moon.

The second channel transmitted the integrated output of a Geiger counter. After the probe had passed through the Van Allen belts, and there was no more information to be measured, the input to this channel was switched to measure the output of the power amplifier of the transmitter.

Channel 3 was a method for transmitting the number of counts from a second Geiger-Mueller tube. Three signals were superimposed as shown in figure 7.1. The highest frequency output from the counter switched every 2^9 counts, the next every 2^{13} counts, and the last every 2^{17} counts. These three signals were added as shown and passed through a low-pass filter. As the counting rate increased, the high-frequency component would be filtered out and, hence, the number of counts would be scaled by a factor of 2^{13} rather than 2^9. As the number of counts per second increased even further, the middle frequency would also be filtered and the scale factor would be 2^{17}. The most significant digits of the count were thereby always transmitted, while the bandwidth necessary to do this was kept relatively constant. This is an example of a self-adaptive data-processing system.

The data signals from the three channels were applied to the input of three VCO's, each centered at a different frequency to keep the outputs from overlapping in frequency. The signals were actually filtered first, effectively changing the square waves to exponentially increasing and decreasing waveforms as shown in figure 7.2. This filtering was done to prevent the VCO output signal from changing frequency too rapidly so that it could be coherently demodulated with a phase-locked loop. If the frequency input to the demodulating loop were to change too abruptly,

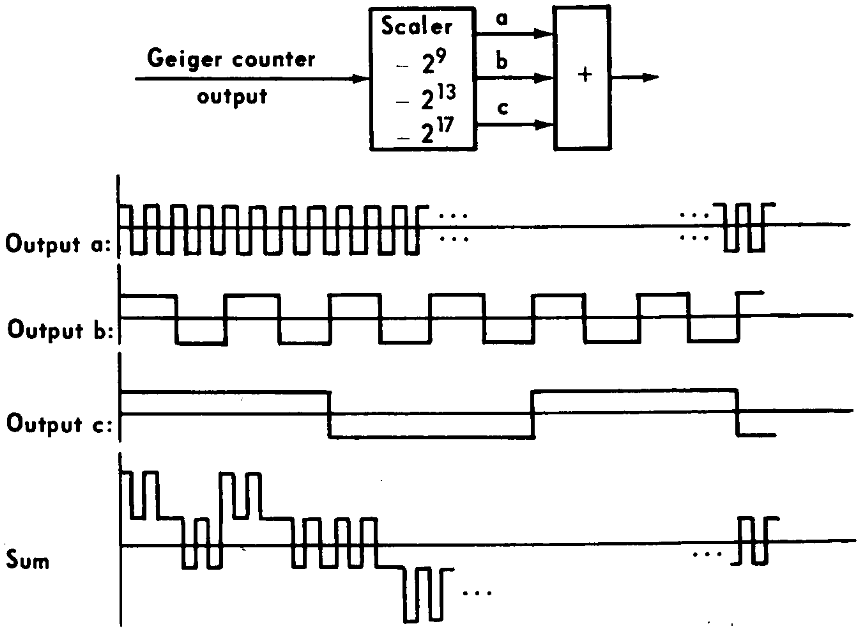


FIGURE 7.1—Channel 3 data signal.

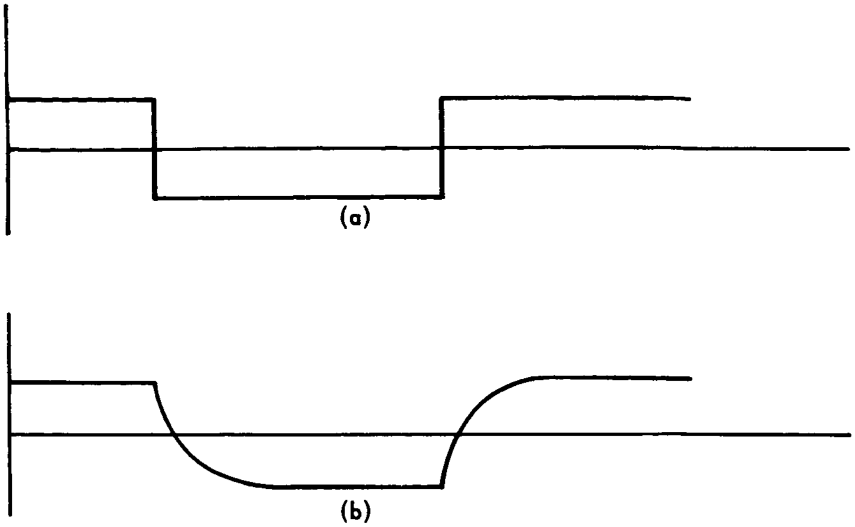


FIGURE 7.2—Data signal conditioning. (a) Prefiltered data signal; (b) filtered data signal.

the loop would go out of lock, and the information would be lost until it could be again locked up. The three VCO outputs were added and used to phase-modulate the carrier.

The total radiated power was 164 mW, of which approximately 100 mW were in the carrier—14 mW each in channels 1 and 2, and 36 mW in channel 3. The probe antenna was essentially a dipole antenna with a gain of 2.5 dB.

The receiver antenna had an 85-foot diameter with a 40-dB gain. The receiver effective noise temperature was 1630° K. Demodulation was accomplished with phase-locked loops as shown in figure 7.3. The loop-noise bandwidths were carrier, 20 cps; channel 1, 4 cps; channel 2, 4 cps; channel 3, 8 cps.

Note that the carrier demodulation is accomplished by simply tracking the unmodulated carrier and taking as the partially demodulated signal the product of the tracked carrier with the received signal. This achieves the desired result only because the carrier is phase modulated with a small index of modulation. That is, let $x_1(t)+x_2(t)+x_3(t)=\phi(t)$, where $x_i(t)$ is the frequency-modulated signal from the ith channel. Then the output of the phase modulator is

$$\begin{aligned} y(t) &= \sqrt{2}A \sin [\omega_c t + \Delta\theta\phi(t)] \\ &= \sqrt{2}A \cos [\Delta\theta\phi(t)] \sin \omega_c t + \sqrt{2}A \sin [\Delta\theta\phi(t)] \cos \omega_c t \end{aligned} \qquad (7.1)$$

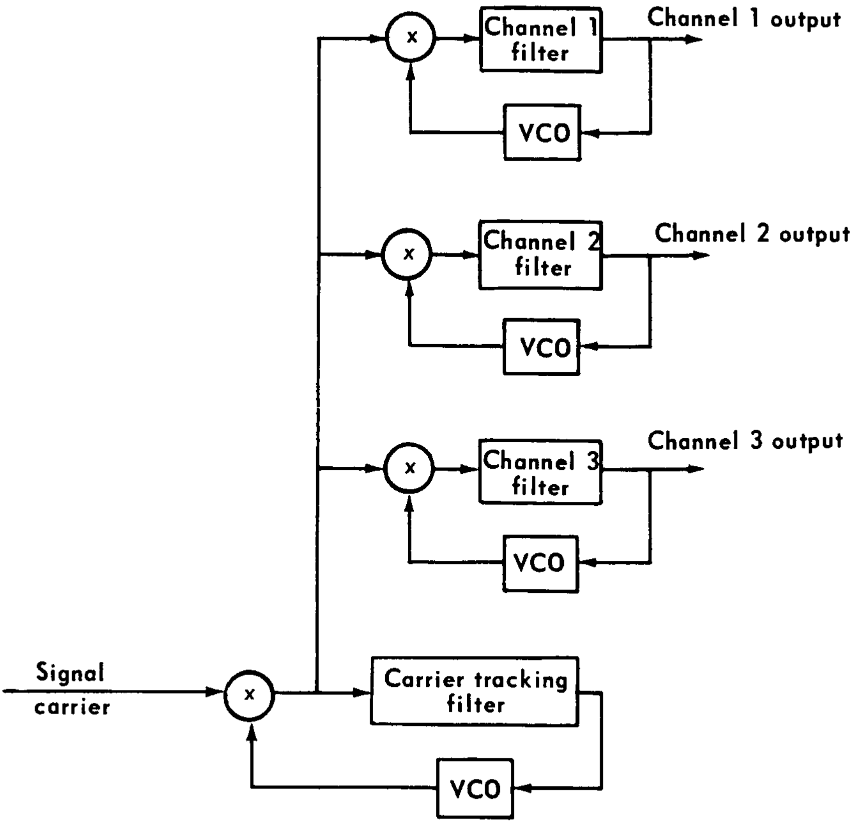


FIGURE 7.3—Pioneer receiver demodulator.

If $\Delta\theta$ is small enough, then

$$\cos\,[\Delta\theta\phi(t)]\approx 1$$

$$\sin\,[\Delta\theta\phi(t)]\approx\Delta\theta\phi(t)$$

and

$$y(t)\approx\sqrt{2}A\,\sin\,\omega_ct+\sqrt{2}A\Delta\theta\phi(t)\,\cos\,\omega_ct \tag{7.2}$$

The carrier loop tracks the unmodulated component at the frequency $f_c=\omega_c/2\pi$. The output of the carrier loop VCO is therefore approximately $\sqrt{2}\,\cos\,\omega_ct$, and, neglecting the double frequency components, the product of this with the signal yields

$$A\Delta\theta\phi(t)=A\Delta\theta[x_1(t)+x_2(t)+x_3(t)] \tag{7.3}$$

which is the desired frequency-multiplexed frequency-modulated signal.

Each of the three loops is designed to track only one of these three signals; the other two are eliminated by the tracking filter. That the FM signal is then demodulated follows from the discussion in "Demodulation of Angle Modulated Signals" in chapter 3.

The Pioneer IV telecommunications system was then essentially a frequency-multiplexed FM system which was coherently demodulated using phase-locked loops. No direct effort was made to remove the redundancy from the data signals, though, clearly, all three channels produced very redundant information. Some effort was expended, however, to make use of the signal redundancy by superimposing several signals. This was, in some sense, a method of data compaction. In addition, the third channel did incorporate a simple but effective method of adapting to the data.

THE MARINER II COMMUNICATIONS SYSTEM

The Mariner II telemetry system provides an interesting contrast to the Pioneer system. First, it was designed for a communication distance of up to 40 million miles as opposed to the ¼- to ½-million-mile required range of the lunar probe. In addition, many more data sources were on board the spacecraft. These data sources can be divided into engineering and scientific sources. The engineering measurements included: measurements of the battery voltage; the solar-panel voltage; the earth-sensor temperature; roll, pitch, and yaw gyros; sun sensors; propellant tank pressure and temperature; thermal-shield temperatures, etc. Some of the scientific experiments which had to be monitored included a microwave radiometer experiment, an infrared radiometer experiment, a magnetometer experiment, charged particle flux experiments, solar plasma experiments, and a micrometeorite experiment.

To complicate the situation were the following two factors: (1) not all measurements were needed at all times; and (2) the telemetry capacity obviously decreased as the distance from the earth increased. It clearly would not be efficient to operate as if the worst condition were in effect; that is, as if the spacecraft were at its maximum design distance from the earth and all measurements were to be transmitted. Fortunately, many of the engineering measurements are most significant during the early powered stages of the flight and during the midcourse maneuver, while the telemetry capacity is relatively large and the scientific measurements are not of interest. The scientific measurements become most important during the fairly brief period of planetary encounter when the spacecraft has approached its maximum design distance from the earth. For these reasons, the telemetry system was designed to operate in three successive modes: the launch mode, the cruise mode (after earth acquisition), and the planetary encounter mode (during which the scientific data were gathered and transmitted). Switching from one mode to another was effected by a command from the ground.' Each telemetry mode had associated with it a certain subset of the total measurements which were to be monitored.

The various data sources were time multiplexed, each of the signals being observed continually for a fixed length of time T_W. A clock generating $R_W=1/T_W$ pulses per second switched the different measurements into the output. Since some measurements needed to be observed more frequently than others, a process of commutation evolved so that it was possible to sample some sources every 10th time, some every 100th time, and some only every 1000th time. The technique for accomplishing this is indicated in figure 7.4. The switches marked S indicate some of the various possible mode switches. The scientific data were transmitted in the encounter mode. The 10-input, 1-output boxes indicate commutators which read the input from the ith switch until it receives a pulse and then switches to the $i+1$st switch. After it reaches the 10th position, it starts over again at the 1st. At the output the information is sampled once every T_W seconds and converted to binary PCM form. Each sample is represented by words of seven bits each. The bit clock therefore generates $R_B=7R_W$ pulses per second; the bit interval is $T_B=1/R_B$ seconds.

The input to the modulator is a sequence of PCM binary bits. At the output of the modulator a 1 is represented by the signal $x_1(t)=\sqrt{2}A$ sin $\omega_s t$ and a zero by $x_2(t)=-\sqrt{2}A$ sin $\omega_s t$. Since the time signals represent equal power, and since

$$\rho_{12}=\frac{\int_0^T x_1(t)x_2(t)\ dt}{[\int_0^T x_1^2(t)\ dt \int_0^T x_2^2(t)\ dt]^{1/2}}=-1 \tag{7.4}$$

this corresponds to a two-level biorthogonal code. The associated bit-error probabilities are as shown in figure 4.11, for $m=\log_2 M=1$.

A convenient alternative way of representing the modulator output is by the waveform

$$\sqrt{2}A \cos [\omega_s t+\pi/2m(t)] \tag{7.5}$$

where $m(t)=+1$ or -1 for $nT_B<t<(n+1)T_B$ depending upon whether the bit is a 1 or a zero. This is therefore an example of discrete phase-shift keying (PSK).

This PSK signal is, in turn, used to phase-modulate the 960-Mc/sec carrier. The same carrier demodulators are used that were described in "The Pioneer Communications System" in this chapter. Again, because of the nonlinear aspects of coherent PM demodulation, it is

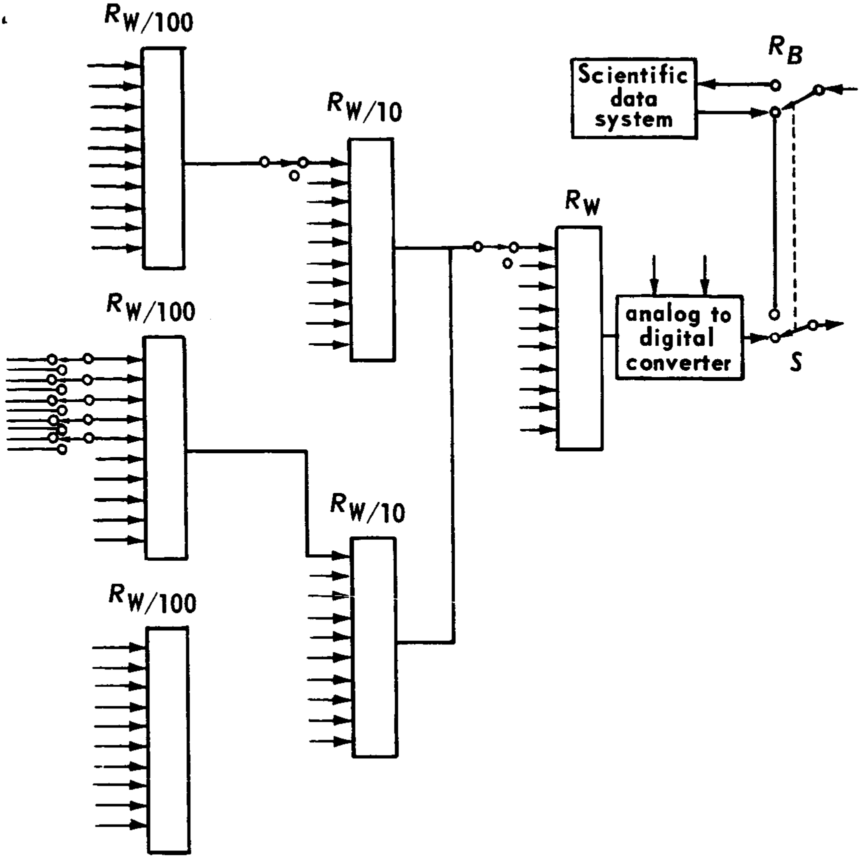


FIGURE 7.4—Telemetry commutator and mode-switching mechanization.

required that about 50 percent of the power be left in the carrier; that is, that the modulation index be kept relatively small.

In practice, instead of the signal in the expression (7.5), the signal

$$\sqrt{2}A \cos [\omega_s t+\theta m(t)] \tag{7.6}$$

is used where θ is somewhat less than $\pi/2$. This signal is used because coherent demodulation demands knowledge of the phase and frequency of the unmodulated subcarrier and

$$\begin{aligned}\sqrt{2}A \cos [\omega_s t+\theta m(t)] &= \sqrt{2}A \cos [\theta m(t)] \cos \omega_s t-\sqrt{2}A \sin [\theta m(t)] \sin \omega_s t\\ &= \sqrt{2}A \cos \theta \cos \omega_s t-\sqrt{2}Am(t) \sin \theta \sin \omega_s t\\ &= \sqrt{2}A \cos \theta \cos \omega_s t+\sqrt{2}A \sin \theta \cos \left[\omega_s t+\frac{\pi}{2}m(t)\right] \end{aligned} \tag{7.7}$$

Thus, the signal $\sqrt{2}A$ cos $[\omega_s t+\theta m(t)]$ may be considered to be the sum of an unmodulated subcarrier of amplitude $\sqrt{2}A$ cos θ and a ±90° phase-modulated sinusoid with amplitude $\sqrt{2}A$ sin θ. The subcarrier is demodulated by tracking the unmodulated portion with a phase-locked loop and taking the product of the total signal with the retrieved pure subcarrier, as shown in figure 7.5. The product

$$\sqrt{2} \sin \omega_s t[\sqrt{2}A \cos \theta \cos \omega_s t - \sqrt{2}Am(t) \sin \theta \sin \omega_s t]$$

becomes, after filtering out the double frequency terms

$$-Am(t) \sin \theta$$

which, except for the insignificant (constant) change of sign and gain, is the original data signal.

Note that the optimum demodulation scheme as discussed in chapter 4 is used to convert the signal $-Am(t)$ sin θ to a sequence of binary bits. This, of course, demands the knowledge of the instants of time when one bit ends and the next one begins. This bit synchronization as well as

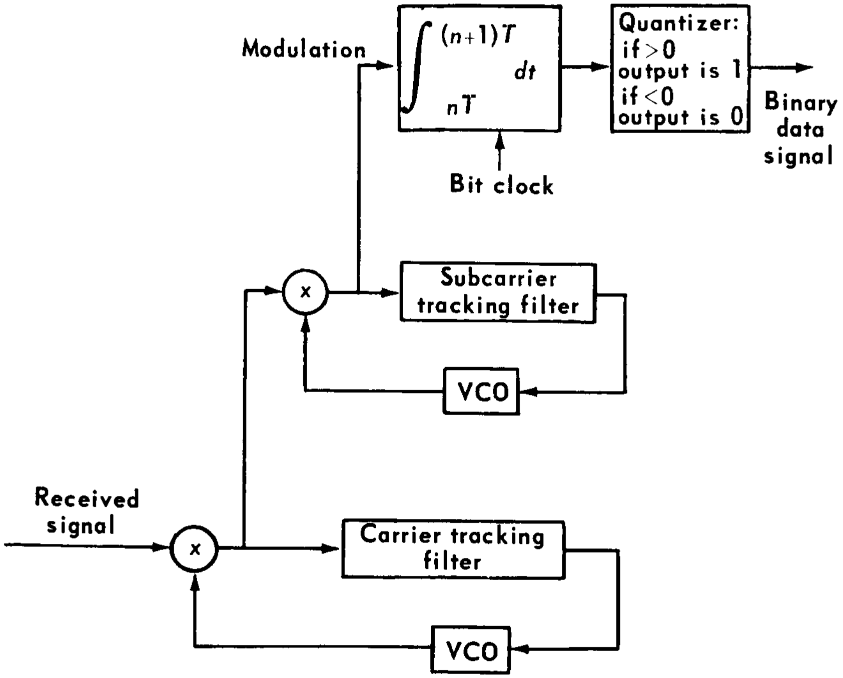


FIGURE 7.5—The Mariner II telemetry demodulator.

word and frame synchronization must be provided as discussed in "Synchronization" in chapter 4. The technique by which this information was obtained in the Mariner II system is entirely analogous to the PR ranging system discussed in "Ranging With Pseudo-Random Sequences" in chapter 6. As pointed out there, ranging and synchronization are strongly related problems; in this case essentially the same solution was used for synchronization as is often used for ranging. A PR sequence containing 63 bits with period $T_B/9$ was transmitted. Since $63(T_B/9)=7T_B=T_W$, the period of the PR sequence was equal to the word interval. By identifying a particular phase of the PR sequence with the beginning of a word, word synchronization, and *a fortiori*, bit synchronization, is obtained as soon as the phase of the PR sequence is detected. The PR-bit period was chosen to be $T_B/9$ rather than T_B in order to increase the resolution; that is, in order to decrease the region of uncertainty Δt (cf. "Ranging With Pseudo-Random Sequences" in ch. 6).

Frame synchronization, the knowledge of that instant of time in which the complete cycle of data observations begins again, was provided by making the first word of each frame consist of 1's only and prohibiting the all-1's word from occurring elsewhere. (If the data signal were truly all 1's, it would be replaced by the signal 1 1 1 1 1 1 0.) Consequently, the occurrence of the all-1's word indicated the beginning of a new frame.

In summary, then, the Mariner II telecommunications system might be classified as a PCM/PSK/PM system. The data are time multiplexed, sampled, and converted to binary PCM. The binary bits are used to PSK-modulate a subcarrier which in turn phase-modulates a carrier. The purpose of a subcarrier is to move the data spectrum away from the zero frequency range so that, when this signal is used to modulate the carrier, the receiver carrier loop is able to track the carrier without interference from the modulation. Approximately 50 percent of the total power must be left in the carrier. This is not necessarily a limitation, however, since unmodulated carrier power is essential both for locking up the receiver loop and for tracking the vehicle. The Mariner II transmitting antenna was parabolic with a 4-foot diameter. The radiated power was 3 watts; the telemetry rate was 33⅓ bits/sec in the launch mode and 8⅓ bits/sec in the other two modes of operation.

The command link (ground to spacecraft) was entirely similar to the telecommunications link. Several important differences should be mentioned, however. First, while it is important to have reliable communication between the spacecraft and ground, it is imperative to have extremely high reliability in the other direction. Commands from the earth are used to alter the mode of operation and even the trajectory of the spacecraft. An error in the reception of a command could easily defeat the purpose of the entire mission. Although much more power (up to 10 kW) is radiated from the ground than from the spacecraft, the ground-to-vehicle link is not significantly better than the vehicle-to-ground link. This lack of improvement results from the vehicle-receiver effective noise temperature (about 5800° K) being considerably higher than that of the ground receiver. In addition, the vehicle-receiver antenna gain is less than its transmitter antenna gain. This latter factor is a result of the requirement that commands must be received regardless of the attitude of the vehicle so that the receiver antenna must be omnidirectional and therefore ideally has a gain of 0 dB. Should something go wrong with the vehicle attitude control, for example, the rest of the mission might be a total failure were it not possible to communicate with the probe regardless of its orientation. The reliability of the ground-to-vehicle communication is increased, however, by transmitting the data at a very slow bit rate. As seen in figure 4.11, the error probability is a rapidly decreasing function of the parameter ST_B/N_0, where T_B is the time spent per bit. Since relatively few commands need to be transmitted, T_B can be made quite long (1 second, in this case), and the bit-error probability, satisfactorily small.

THE DEEP SPACE INSTRUMENTATION FACILITY

The Deep Space Instrumentation Facility (DSIF) consists of three transmitting-receiving sites—Goldstone, Calif.; Woomera, Australia; and Johannesburg, South Africa. Since the facilities at the three sites are comparable, only the station at Goldstone is discussed here.

The Goldstone station, as now equipped, operates at both L-band (about 950 Mc/sec) and S-band (about 2300 Mc/sec). Most systems are now designed to operate in the S-band region due to the increased antenna gain there (cf. "Antenna Gain" in ch. 2). The receiving and transmitting antennas are both 85-foot paraboloids (one with an az-el mount, the other with a HA–DEC mount), having an effective area of about 67 percent of their actual area. They are able to track to within an angular accuracy of approximately 0.02°. The receiver amplifier is a helium-cooled ruby maser and the overall system noise temperature is about 33° K. (This may be broken down roughly as 10° K sky or background noise, 9° K maser noise, and 14° K noise due to various system losses.) The maser gain is about 40 dB over a bandwidth of 12 megacycles. The transmitter is capable of delivering up to 100 kilowatts of continuous power.

THE DIRECTION OF FUTURE SYSTEMS

One improvement at the Goldstone DSIF which is already in progress is the construction of a 210-foot parabolic antenna which will still have tolerances satisfactory for S-band operation. Some consideration is being given to the further increase of power, perhaps by as much as a factor of 5, but this procedure clearly has its limitations. Further significant decreases in the receiver noise temperature below the present 33° K are not anticipated.

So far as the communication system itself is concerned, the most probable direction will be toward the use of multilevel orthogonal or biorthogonal codes. As seen in figures 4.10 and 4.11, the advantages of going to higher level codes can be significant. The factors limiting their use up to now have been the inherently more complex synchronization problem and the somewhat increased equipment complexity. Neither of these problems now seems to be severe.

Some improvement in the now quite limited data compaction processes is also likely to be incorporated into future systems. The onboard equipment limitation is still not insignificant in this case, however, and will continue to restrict the improvements which can be made in this direction.

Bibliography

CHAPTER 1

DAVENPORT, WILBUR B. JR.; AND ROOT, WILLIAM L.: An Introduction to the Theory of Random Signals and Noise. McGraw-Hill Book Co., Inc., 1958.

RICE, STEVEN O.: Mathematical Analysis of Random Noise. Bell System Tech. J., vol. 23, no. 3, 1944, pp. 282–332.

RICE, STEVEN O.: Mathematical Analysis of Random Noise. Bell System Tech. J., vol. 24, no. 1, Jan. 1945, pp. 46–156.

CHAPTER 2

BLACKWELL, LAWRENCE A.; AND KOTZEBUE KENNETH L.: Semiconductor-Diode Parametric Amplifiers. Prentice-Hall, Inc., 1961.

HOGG, CHRISTOPHER A., AND SUCSY, LAWRENCE G.: Masers and Lasers. Maser & Laser Assoc., Cambridge, Mass., 1962.

FRADIN, C. Z.: Microwave Antennas. Pergamon Press, 1961.

HANSEN, R. D., ED.: Microwave Scanning Antennas. Vol. I, Academic Press, 1964.

PIERCE, JOHN R.: Electrons, Waves and Messages. Hanover House, 1956.

CHAPTER 3

BLACK, HAROLD S.: Modulation Theory. D. Van Nostrand Co., Inc., 1953.

SCHWARTZ, MISCHA: Information Transmission, Modulation and Noise. McGraw-Hill Book Co., Inc., 1959.

JAFFE, R.; AND RECHTIN, E.: Design and Performance of Phase-Lock Circuits Capable of Near-Optimum Performance Over a Wide Range of Input Signal and Noise Levels. IRE Trans. on Information Theory, Mar. 1955, pp. 66–76.

CHAPTER 4

FANO, ROBERT M.: Transmission of Information. MIT Press and John Wiley & Sons, Inc., 1961.

GOLOMB, SOLOMON W., ED.: Digital Communications With Space Applications. Prentice-Hall, Inc., 1964.

CHAPTER 5

ELIAS, PETER: Predictive Coding. IRE Trans. on Information Theory, Mar. 1955, pp. 16–33.

HUFFMAN, D. A.: A Method for the Construction of Minimum Redundancy Codes. Proc. IRE, vol. 40, Sept. 1952, p. 1098.

CHAPTER 6

GOLOMB, SOLOMON W., ED.: Digital Communications With Space Applications. Prentice-Hall, Inc., 1964.

CHAPTER 7

MARTIN, BENN D.: The Pioneer IV Lunar Probe: A Minimum Power FM/PM System Design. JPL Tech. Rep. no. 32–225, Mar. 1962.

RIDDLE, F. M.; ET AL.: JPL Contributions to the 1962 National Telemetering Conference. JPL Tech. Memo. no. 33-88, May 1962.

☆ U. S. GOVERNMENT PRINTING OFFICE : 1967 O - 217-912

> *"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."*
>
> —NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Technical information generated in connection with a NASA contract or grant and released under NASA auspices.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

TECHNOLOGY UTILIZATION PUBLICATIONS: Information on technology used by NASA that may be of particular interest in commercial and other nonaerospace applications. Publications include Tech Briefs; Technology Utilization Reports and Notes; and Technology Surveys.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION DIVISION
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Washington, D.C. 20546